Spring 5-15-2017

Model and World: Generalizing the Ontic Conception of Scientific Explanation

Mark Povich
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Philosophy-Neuroscience-Psychology Program

Dissertation Examination Committee:
Carl F. Craver, Chair
Stuart Glennan
John Heil
Ron Mallon
Anya Plutynski

Model and World: Generalizing the Ontic Conception of Scientific Explanation

by
Mark Adam Povich

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2017
St. Louis, Missouri

Table of Contents

Acknowledgments

I would like to thank Andre Ariew, Mike Dacey, Dylan Doherty, Stuart Glennan, John Heil, Eric Hochstein, Philippe Huneman, Daniel Kostic, Ron Mallon, Joe McCaffrey, Christiane Merritt, Gualtiero Piccinini, Anya Plutynski, Rick Shang, Kate Shrumm, Julia Staffel, Catherine Stinson, Dan Weiskopf, David Wright, Tom Wysocki, and audience members at the 2014 and 2016 Philosophy of Science Association conferences and the 2014 SLAPSA conference for feedback on various parts of this work. I would also like to thank my department secretaries Mindy Danner, Sue McKinney, Kimberly Mount, and Dee Stewart. Finally, for financial, emotional, and philosophical support and friendship, I would like to thank my advisor, Carl Craver.

Mark Povich

Washington University in St. Louis

May 2017

ABSTRACT OF THE DISSERTATION

Model and World: Generalizing the Ontic Conception of Scientific Explanation

by

Mark Adam Povich

Doctor of Philosophy in Philosophy-Neuroscience-Psychology

Washington University in St. Louis, 2017

Professor Carl F. Craver, Chair

Model and World defends a theory of scientific explanation that I call the "Generalized Ontic Conception" (**GOC**), according to which a model explains when and only when it provides (approximately) veridical information about the ontic structures on which the explanandum phenomenon depends. Causal and mechanistic explanations are species of **GOC** in which the ontic structures on which the explanandum phenomenon depends are causes and mechanisms, respectively, and the kinds of dependence involved are causal and constitutive/mechanistic, respectively. The kind of dependence relation about which information is provided determines the species of the explanation. This provides an intuitive typology of explanations and opens the possibility for non-causal, non-mechanistic explanations that provide information about non-causal, non-mechanistic kinds of dependence (Pincock 2015; Povich forthcoming a). What unites all these forms of explanation is that, by providing information about the ontic structures on which the explanandum phenomenon depends, they all can answer what-if-things-had-been-different questions (w-questions) about the explanandum phenomenon. This is what makes causal explanations, mechanistic explanations, and non-causal, non-mechanistic explanations all *explanations*.

Furthermore, **GOC** is a generalized ontic conception of scientific explanation (Salmon 1984, 1989; Craver 2014). It is consistent with Craver's claim that (2014), according to the ontic conception, commitments to ontic structures (like causes or mechanisms) are required to demarcate explanation from other scientific achievements. **GOC** demarcates explanatory from non-explanatory models in terms of ontic structures. For example, the distinction between explanatory and phenomenal models is cashed out in terms of the ontic structures about which information is conveyed: A phenomenal model provides information about the explanandum phenomenon, but not the ontic structures on which it depends. **GOC** is generalized because it says that commitments to more of the ontic than just the causal-mechanical – the traditional focus of the ontic conception – are required adequately to achieve this demarcation; attention to all ontic structures on which the explanandum depends is required.

The relation between model and world required for explanation is elaborated in terms of information rather than mapping, reference, description, or similarity (Craver and Kaplan 2011; Kaplan 2011; Weisberg 2013). The latter concepts prove too strong, so will not count models as explanatory that in fact are. Take Kaplan and Craver's (2011) model-to-mechanism-mapping (**3M**) principle. According to **3M**, the variables in a mechanistic explanatory model must map to specific structural components and causal interactions of the explanandum phenomenon's mechanism. However, you can mechanistically explain without referring to the explanandum's mechanism or its components and their activities, for example, by describing what the mechanism is *not* like. This is a way of constraining or conveying information about a mechanism without actually mapping to, referring to, describing, representing, or being similar to it.

Chapter 1. Introduction: Brief Historical Positioning

Salmon (1989) began his authoritative history of scientific explanation with Hempel and Oppenheim's "Studies in the Logic of Explanation," published in 1948. If we follow Salmon in so dating the birth of the philosophical study of scientific explanation, then we are now nearing the end of its seventh decade. What lessons can we draw from the intervening three decades between Salmon's study and today? The central idea of this dissertation is that a wrong turn was made during the third decade, solidified in the fourth, and down which we have traveled to this day. In this introductory chapter, I will explain why this turn was made and how to get back on the right path, which, it turns out, is not so far away.

According to Hempel and Oppenheim's (1948) account, which was to become known as the deductive-nomological (DN) model, an explanation is an argument with descriptions of at least one law of nature and antecedent conditions as premises and a description of the explanandum phenomenon, the thing to be explained, as the conclusion. On this view, to explain is to show that the explanandum phenomenon is predictable on the basis of at least one law of nature and certain specific antecedent and boundary conditions.

The DN model soon became the received theory of scientific explanation, one of two "grand traditions" (Salmon 1990), although it always had critics. An early critic was Scriven (1958), who was to start the second "grand tradition," according to which explaining a phenomenon is identifying its cause(s). The first tradition, according to Salmon (1989; 1990), lives on in unificationism (Friedman, 1974; Kitcher 1989). The basic idea behind unificationism is that explaining an event is deriving a description of it from an argument pattern that can be used to derive descriptions of many different phenomena. This unifies the explanandum

phenomenon with the other phenomena that are derivable by the same argument pattern, thereby showing it to be part of a more general pattern of phenomena. The second tradition lives on in mechanistic approaches (Bechtel and Abrahamsen 2005; Craver 2006; Glennan 2002; Kaplan 2011; Machamer, Darden, and Craver 2000) and causal/difference-making approaches (Strevens 2008; Woodward 2003; Woodward and Hitchcock 2003) to explanation.

In the intervening three decades since Salmon published *Four Decades of Scientific Explanation,* I think it is safe to say that more philosophers have signed on to causal-mechanical accounts of explanation than to unificationist or DN accounts (though see Sansom and Shields [2016] and Schweder [2005]). One consequence of this is that causal-mechanical accounts have become the new main targets of philosophical critique (Batterman 2002a; 2002b; Bokulich 2011; Huneman 2010; Reutlinger 2014; 2016; Rice 2012; 2013; Saatsi and Pexton 2013; Woodward 2013). I think there is something right in these philosophers' critiques, but what exactly are causal-mechanical accounts leaving out? To address this question, I want to briefly consider the historical motivations for causal-mechanical accounts of explanation.

According to Salmon (1989, 116), growing realization of the central role of causation in explanation came during the third decade, and was further developed in the fourth decade. The following are two of the influential counterexamples to the DN model that motivated this growing realization (Salmon 1989, 46).

The length of a flag pole's shadow can be derived from the height of the pole and the angle of elevation of the sun (Bromberger 1966). Therefore, according to the DN model, the length of a flag pole's shadow is explained by the height of the pole and the angle of elevation of the sun. But this derivation is symmetric. That is, one can also derive the height of the flag pole

from the length of its shadow and the angle of elevation of the sun. Therefore, according to the DN model, the height of the flag pole is explained by the length of its shadow and the angle of elevation of the sun. But that does not seem right. Here it is plausible that the real explanatory work is done by causation: the derivation of the length of the pole's shadow counts as explanatory because that derivation, but not the reverse derivation, tracks causes.

Another counterexample concerns common causes. A drop in atmospheric pressure causes the oncoming storm and the barometer to dip (Salmon 1989, 47). We can use the dipping barometer to predict an oncoming storm. Since the DN model ties explanation to prediction, the dipping barometer, therefore, explains the oncoming storm. Again, that does not seem right. The drop in atmospheric pressure explains both the oncoming storm and the dip in the barometer. Causation, again, does the explanatory work.

The lesson that was drawn by many philosophers from these counterexamples to the DN model was that explanation in general is a matter of identifying causes. We can now see in hindsight, I argue, that this was the wrong general lesson to draw. It is correct that explanation in *these* cases is a matter of identifying causes. There is a lesson that can be drawn about explanation in general, but it can only be seen in hindsight: explanation in general is a matter of identifying ontic relations of dependence. This lesson is actually a version of Salmon's ontic conception of scientific explanation.

Salmon (1984; 1989) distinguished between epistemic, modal, and ontic conceptions of explanation. These are conceptions of what a scientific explanation aims to show of the explanandum phenomenon: that it is expected to occur, that it had to occur, that it fits “into a discernible pattern,” respectively (1984, 121). For Salmon, the “discernible pattern” into which

the explanandum phenomenon is fit is structured by causal processes, causal interactions, and causal laws (Ibid., 132). "[W]e explain," wrote Salmon, "by providing information about these patterns that reveals how the explanandum-events fit in" (1989, 121).

The right lesson to be drawn from the DN counterexamples for the theory of explanation in general is that explanation is not about nomic expectability, but about fitting the explanandum into "discernible patterns," "relationships that exist in the world" (1984, 121). But, I argue, this should not be construed solely in term of causation. It is a mistake to equate the ontic conception with the causal-mechanical account of explanation, as both critics and advocates of Salmon assume. Salmon actually did not think causation was essential to the ontic conception:

> It could fairly be said, I believe, that mechanistic explanations tell us how the world works. These explanations are local in the sense that they show us how particular occurrences come about; they explain particular phenomena in terms of collections of particular causal processes and interactions – or, perhaps, in terms of *noncausal* mechanisms, if there are such things. (1989, 184; my emphasis)

For Salmon, what was essential to the ontic conception was that, "the explanation of events consists of fitting them into the patterns that exist in the objective world" (1989, 121). We can and should hold on to the ontic conception while accepting many of the criticisms and limitations of causal explanation provided in the three decades since *Four Decades*. There are *noncausal* dependence relations in which an explanandum phenomenon can stand to other worldly items (examples are discussed in Chapter 4). I call the resulting view the Generalized Ontic Conception **(GOC)** of scientific explanation.

According to the **GOC**, a model explains when and only when it provides (approximately)[1] veridical information about the ontic structures[2] on which the explanandum

[1] Idealization will be dealt with in Chapter 4.

[2] The term "structure" can have technical senses, e.g., in the philosophy of mathematics. Here I use it as a catch-all for worldly properties, relations, events, objects, and so forth, possibly even structures in the mathematical sense (see Chapter 8).

phenomenon depends. (Note that Salmon said there were two ways to formulate an ontic conception: explanations are certain ontic structures or explanations are descriptions of them [1989, 86].) Causal and mechanistic explanations are species of **GOC** in which the ontic structures on which the explanandum phenomenon depends are causes and mechanisms, respectively, and the kinds of dependence involved are causal and constitutive/mechanistic, respectively. The kind of dependence relation about which information is provided determines the species of the explanation. This provides an intuitive typology of explanations and opens the possibility for non-causal, non-mechanistic explanations that provide information about non-causal, non-mechanistic kinds of dependence (Pincock 2015; Povich forthcoming a; see also Lowe [2013] on kinds of metaphysical dependence). What unites all these forms of explanation is that, by providing information about the ontic structures on which the explanandum phenomenon depends, they all can answer what-if-things-had-been-different questions (w-questions) about the explanandum phenomenon. This is what makes causal explanations, mechanistic explanations, and non-causal, non-mechanistic explanations all *explanations*.

GOC is also consistent with Craver's (2014) formulation of the ontic conception, according to which commitments to ontic structures (like causes or mechanisms) are required to demarcate explanation from other scientific achievements. **GOC** demarcates explanatory from non-explanatory models in terms of ontic structures. For example, the distinction between explanatory and phenomenal models is cashed out in terms of the ontic structures about which information is conveyed: A phenomenal model provides information about the explanandum phenomenon, but not the ontic structures on which it depends. **GOC** is generalized because it says that commitments to more of the ontic than just the causal-mechanical – the traditional

focus of the ontic conception – are required adequately to achieve this demarcation; attention to all ontic structures on which the explanandum depends is required.

I elaborate the relation between model and world required for explanation in terms of information rather than mapping, reference, description, or similarity (Craver and Kaplan 2011; Kaplan 2011; Weisberg 2013). The latter concepts prove too strong, so will not count models as explanatory that in fact are. Take Kaplan and Craver's (2011) model-to-mechanism-mapping (**3M**) principle. According to **3M**, the variables in a mechanistic explanatory model must map to specific structural components and causal interactions of the explanandum phenomenon's mechanism. However, you can mechanistically explain without referring to the explanandum's mechanism or its components and their activities, for example, by describing what the mechanism is *not* like. This is a way of constraining or conveying information about a mechanism without actually mapping to, referring to, describing, representing, or being similar to it.

Each of the critiques of causal-mechanical accounts of explanation I mentioned earlier (Batterman 2002a; 2002b; Bokulich 2011; Huneman 2010; Reutlinger 2014; 2016; Rice 2012; 2013; Saatsi and Pexton 2013; Woodward 2013) begins to point towards the **GOC** by emphasizing an explanation's ability to answer counterfactual w-questions. However, none of these critiques grapples with the question of what distinguishes explanatorily relevant from irrelevant counterfactuals. It is the ontic component of **GOC** that supplies the asymmetry of explanatory relevance.

Alongside the history of the philosophical study of scientific explanation, philosophers of science became increasingly interested in scientific models (Morgan and Morrison 1999): what

they are (Godfrey-Smith 2009; Weisberg 2013), how we learn about them and use them to learn about the world (Morgan 1999; Weisberg 2013), and how they represent (Contessa 2007; Giere 2004; Suárez 2004). One source of this interest begins with the philosophical analysis of theories (for discussion of the structure of scientific theories, see Craver 2002; Halvorson 2016). On the axiomatic/syntactic view of theories (Hempel 1965; Nagel 1961), which was developed in parallel the DN model of scientific explanation, scientific theories are systems of formal logical structures that are partially interpreted to make empirical claims. According to the semantic view of theories, in contrast, scientific theories are (sets of) models in which the claims made by the theories are true (Suppe 1977; Suppes 1961; 1967; van Fraassen 1980). Critiques of the semantic view of theories have continued to emphasize the importance of models (Craver 2002; Morrison 1999).

However, philosophers primarily interested in the semantics, epistemology, and ontology of models have not usually been interested in what makes models explanatory. Two prominent exceptions are van Fraassen (1980) and Cartwright (1983). According to van Fraassen's (1980) pragmatic theory of explanation, an explanation is an answer to a why-question that consists of an ordered triple of explanandum phenomenon, contrast class, and relevance relation. However, van Fraassen places so few constraints on the relevance relation that in some contexts astral influence can count as a relevance relation, thereby permitting astrological explanations (Kitcher and Salmon 1987; Salmon 1989). According to Cartwright's (1983) simulacrum account of explanation, "to explain a phenomenon is to construct a model which fits the phenomenon into a theory" (17). However, while the simulacrum account appears to have affinities with both the DN model and unificationism, Cartwright provides little guidance on how exactly the account

works. Without more constraints on what a model must do to be explanatory, it seems that any theoretical model that matches the phenomenological behavior of the explanandum is an explanation (161). Cartwright is forthright that her account sheds little light on how causal explanations work (162). **GOC** describes what a model must do to be explanatory and provides a clear taxonomy of kinds of explanation.

In conclusion, Salmon (1990) identified two "grand traditions" in the philosophy of explanation. The first tradition begins with Hempel's DN model and runs through unificationism. The second tradition begins with Scriven and runs through causal-mechanical accounts of explanation. The first tradition adheres to an epistemic conception of explanation while the second adheres to an ontic conception. I argue that it is a mistake to equate the ontic conception with a causal-mechanical account of explanation. My hope is that the **GOC** can provide the material for a new consensus in scientific explanation's eighth decade.

In the following chapters, I explore the explanatory status of models in psychology (Chapters 2 and 3). Finding no non-causal, non-mechanistic explanations there, I turn to models in biology and physics (Chapter 4), where I explicate a new theory of explanation that can accommodate non-causal, non-mechanistic explanation: the **GOC**. I then work on the details of that theory (Chapters 5 and 6) and, with Prof. Carl F. Craver, critique a prominent opposing account (Chapter 7).

In Chapter 2, I consider the question of which psychological models, if any, are mechanistic explanations. That this is a heavily debated question should seem a little strange given that there is rough consensus on the following two claims: 1) A mechanism is an organized collection of entities and activities that produces, underlies, or maintains a phenomenon. 2) A

mechanistic explanation describes or otherwise represents the mechanism producing, underlying, or maintaining the phenomenon to be explained (i.e. the explanandum phenomenon) (Bechtel and Abrahamsen 2005; Craver 2007). If there is a rough consensus on what mechanisms are and that mechanistic explanations represent them, then how is there no consensus on which psychological models are mechanistic explanations? Surely the psychological models that are mechanistic explanations are the models that represent mechanisms. That is true, of course; the trouble arises when determining what exactly that involves. Philosophical disagreement over which psychological models are mechanistic explanations is often disagreement about what it means to represent a mechanism, among other things (Hochstein 2016; Levy 2013). In addition to what it means to represent a mechanism, one's position in this debate arguably depends on a host of other seemingly arcane metaphysical issues, such as the nature of computational and functional properties (Piccinini 2016) and realization (Piccinini and Maley 2014), as well as the relation between models, methodologies, and explanations (Craver 2014; Levy 2013; Zednik 2015). Although I inevitably advocate a position, my primary aim in this chapter is to spell all of these relationships out and canvas the positions that have been taken (or one could take) with respect to mechanistic explanation in psychology, using dynamical systems models and cognitive models (or functional analyses) as examples. This chapter is forthcoming in *The SAGE Handbook of Theoretical Psychology* (Povich forthcoming b).

In Chapter 3, I examine the explanatory status of three specific psychological models. Weiskopf (2011) argues that the models of object recognition and categorization, JIM, SUSTAIN, and ALCOVE, are not mechanistic, yet satisfy the norms of explanation. The norms in question are the ability to manipulate and answer counterfactual questions about the

explanandum phenomenon. In this chapter, I argue that these models are mechanism-sketches. My argument applies recent research using model-based fMRI, a novel neuroimaging method whose significance for current debates on psychological models and mechanistic explanation has yet to be explored. This chapter was published in Povich (2015).

In Chapter 4, I present **GOC** and argue that it better accounts for the explanatoriness of so-called "minimal models" than Batterman and Rice's (2014) account. They argue that minimal models possess explanatory power that cannot be captured by what they call "common features" approaches to explanation, of which **GOC** is a species. Minimal models are explanatory, according to Batterman and Rice, not in virtue of accurately representing relevant features, but in virtue of answering three questions that provide a "story about why large classes of features are irrelevant to the explanandum phenomenon" (356). In this chapter, I argue, first, that a method (the renormalization group) they propose that answers the three questions cannot answer them, at least by itself. Second, I argue that answers to the three questions are unnecessary to account for the explanatoriness of their minimal models. Finally, I argue that a common features account, what I call **GOC** or the generalized ontic conception of explanation, can capture the explanatoriness of minimal models. This chapter is forthcoming in *The British Journal for the Philosophy of Science* (Povich forthcoming a).

In Chapter 5, I examine a potentially threatening implication of **GOC**. I identify an inconsistent triad between any informational theory of explanation like **GOC**, the agent-relativity of information, and a kind of objectivity of explanation. I also note how an informational theory of explanation is implicit in core arguments of mechanists (Piccinini and Craver 2011; Zednik 2015). However, informational theories seem to conflict with some lay and scientific

commonsense judgments and goals of the theory of explanation, because information is relative to the background knowledge of agents (Dretske 1981). Sometimes a model is an explanation *simpliciter*, not just an explanation relative to some particular agent. We would also like a philosophical theory to tell us *when* a model is an explanation *simpliciter*, not just when a model is an explanation relative to some particular agent. I sketch a solution by distinguishing explanation *simpliciter* from explanation-to and relativizing the former to what I call "total scientific background knowledge" (TBSK). This chapter was presented in poster form at the Philosophy of Science Association conference in Atlanta in November 2016 and is currently under review.

In Chapter 6, I discuss an epistemological issue that was gestured at in the previous chapter: the determination of the content of TBSK. Specifically, I defend Bird's (2010; 2014) account of social knowledge (**SK**). **SK** denies that scientific social knowledge supervenes solely on the mental states of individuals. Lackey (2014) objects that **SK** cannot accommodate 1) a knowledge-action principle and 2) the role of group defeaters. I argue that Lackey's knowledge-action principle is ambiguous. On one disambiguation, it is false; on the other, it is true but poses no threat to **SK**. Regarding group defeaters, I argue that there are at least two options available to the defender of **SK**, both taken from literature on individual defeaters and applied to group defeaters. Finally, I argue that Lackey's description of the case of Dr. N. – as a case in which the scientific community does not know but is merely in a position to know – is mistaken. It assumes that Dr. N.'s publication is not scientific knowledge. An analogy to the individual case shows that it is plausible that the scientific community is not merely in a position to know, although its members are. This leaves intact a conception of social knowledge on which it does not supervene

on the mental states of individuals. This chapter has been revised and resubmitted to *Erkenntnis*.

In Chapter 7, co-authored in equal parts research and writing with Carl F. Craver, we critique Lange's account of distinctively mathematical explanation from an ontic perspective. Lange (2013b) uses several compelling examples to argue that certain explanations for natural phenomena appeal primarily to mathematical, rather than natural, facts. In such explanations, the core explanatory facts are modally stronger than facts about causation, regularity, and other natural relations. We show that Lange's account of distinctively mathematical explanation is flawed in that it fails to account for the implicit directionality in each of his examples. This inadequacy is remediable in each case by appeal to ontic facts that account for why the explanation is acceptable in one direction and unacceptable in the other direction. The mathematics involved in these examples cannot play this crucial normative role. While Lange's examples fail to demonstrate the existence of distinctively mathematical explanations, they help to emphasize that many superficially natural scientific explanations rely for their explanatory force on relations of stronger-than-natural necessity. These are not opposing kinds of scientific explanations; they are different aspects of scientific explanation. This chapter has been revised and resubmitted to *Studies in History and Philosophy of Science*.

In Chapter 8, I conclude and discuss future work. There are at least three areas where *Model and World* needs to be further developed. The first is on the explanation/model distinction. I make this distinction in Chapter 2 and Kaplan and Craver (unpublished) make it in their account of norms of explanatory completeness. The second is on model semantics. Although **GOC** does not use terms like “reference” or “mapping,” this does not imply that model semantics has no place in the theory of explanation. The third is on whether the explanatory

power of distinctively mathematical explanations can be accommodated by **GOC**.

Chapter 2. Mechanistic Explanation in Psychology

2.1. Introduction

Among philosophers it is heavily debated which psychological models, if any, are mechanistic explanations. This should seem a little strange given that there is rough[3] consensus on the following two claims: 1) A mechanism is an organized collection of entities and activities that produces, underlies, or maintains a phenomenon. 2) A mechanistic explanation describes or otherwise represents the mechanism producing, underlying, or maintaining the phenomenon to be explained (i.e. the explanandum phenomenon) (Bechtel and Abrahamsen 2005; Craver 2007). If there is a rough consensus on what mechanisms are and that mechanistic explanations represent them, then how is there no consensus on which psychological models are mechanistic explanations? Surely the psychological models that are mechanistic explanations are the models that represent mechanisms. That is true, of course; the trouble arises when determining what exactly that involves. Philosophical disagreement over which psychological models are mechanistic explanations is often disagreement about what it means to represent a mechanism, among other things (Hochstein 2016; Levy 2013). In addition to what it means to represent a mechanism, one's position in this debate arguably depends on a host of other seemingly arcane metaphysical issues, such as the nature of computational and functional properties (Piccinini 2016) and realization (Maley and Piccinini 2014), as well as the relation between models, methodologies, and explanations (Craver 2014; Levy 2013; Zednik 2015). Although I inevitably advocate a position, my primary aim in this chapter is to spell all of these relationships out and canvas the positions that have been taken (or one could take) with respect to mechanistic explanation in psychology, using dynamical systems models and cognitive models (or functional

[3] See Section 2.2.1 below.

analyses[4]) as examples.

In Section 2.2, I lay out the basic conceptual toolkit of and motivation for a mechanistic account of explanation, including only recent historical development (for a more extensive history of mechanistic philosophy, see Chapters 2 and 3 of Glennan and Illari [forthcoming]). In Section 2.3, I analyze more closely the question of what it takes for an explanation to be mechanistic. Taking center stage is an increasingly common distinction between mechanistic explanations, on the one hand, and their representational form (including the methodologies used to construct those representations), on the other (Craver 2014; Hochstein 2016; Levy 2013; Zednik 2011, 2015). I illustrate the way this distinction is used with regards to dynamical systems models, which dynamicists have claimed to be non-mechanistic explanations. A similar dialectic occurs with respect to the mechanistic status of functional analyses. I take this up in Section 2.4, where I examine the issue of the autonomy of psychology and the relation between functionalism and mechanistic explanation. In Section 2.5, I compare the previous concepts and distinctions with the long-standing, though changing, distinction between ontic and epistemic conceptions of scientific explanation.

2.2. Mechanisms and Mechanistic Explanation

I first briefly gesture at an ontology of mechanisms, laying out only the bare commitments required to establish a broad concept of mechanisms and mechanistic levels. Then, I motivate a mechanistic account of explanation, and make two normative distinctions: between mechanism schemata and mechanism sketches and between how-possibly and how-actually models. I also contrast both of those distinctions with phenomenal models.

[4] Not all functional analyses are cognitive models, but all cognitive models are functional analyses, at least as I will use those terms here. See Section 2.4.

2.2.1 Mechanisms

While it is true that there is rough consensus that mechanisms contain entities and activities, or simply active entities, spatiotemporally organized to give rise to a behavior or property of the whole mechanism[5], there is disagreement over the specific metaphysics of mechanisms (see Illari and Williamson [2012] for a discussion of this disagreement and a recommendation of a broad construal of mechanisms, similar to mine, that applies across sciences). I do not wish to get involved in this debate here. I will assume a permissive[6] concept of mechanism as any collection of entities, also broadly construed, whose collective, organized activity gives rise to the behavior or property of a whole in context (also see Levy [2014, 9] on what he calls the 'narrow picture' and the 'broad picture' of mechanisms). The entities in a mechanism need not be neatly localizable or contained within well-defined boundaries. An entity could be any set of structural properties that is robustly detectable (Piccinini and Craver 2011, 296).

Though permissive, this concept of mechanism is not trivial because it does not make every system – not even every *causal* system – a mechanism. Mechanisms contrast with aggregates, which lack the requisite organization. The parts of mechanisms have spatiotemporal properties, and stand in organizational and causal relations to one another, that are explanatorily relevant to the behavior of the mechanism as a whole. As such, mechanisms are more than the sums of their parts: their behavior depends on the spatial, temporal, and causal organization of

[5] I will speak of the "behavior" or "property" of a whole mechanism as that for which the mechanism is responsible, but there is also disagreement about how metaphysically to characterize the phenomenon produced by a mechanism (Kaiser and Krickel forthcoming).

[6] The more restrictive one makes the concept of mechanism (for example, by requiring modularity and stability [Woodward 2013] or localizability [Weiskopf 2011]), the correspondingly rarer mechanistic explanations will be. When presented with a putatively non-mechanistic explanation, one should always ask what concept of mechanism is in the background.

their parts. Aggregates, in contrast, are systems – even causal systems – whose behavior does not depend on the spatial, temporal, and causal organization of their parts. As such, a property of an aggregate is literally a sum of the properties of its parts. The concentration of a fluid, for example, is an aggregation of particles. Aggregates have properties that do not change when their parts are reorganized, because in true aggregates, spatial, temporal, and causal organization is irrelevant (Wimsatt 1997; Povich and Craver forthcoming).

Mechanisms are often organized hierarchically into levels (Craver 2015; Povich and Craver forthcoming). The components of mechanisms can themselves be composed of organized components that are responsible for their activity. Similarly, a mechanism may compose an active entity that is itself a component in a larger mechanism. The term 'mechanistic levels' refers to this embedded, hierarchical organization of mechanisms.

Mechanistic levels contrast with another prominent use of the term “levels” in psychology: Marr's levels (Marr 1982). Marr's levels are best understood as levels of description, abstraction, or realization. The computational level, algorithmic level, and implementational level arguably do not stand in causal or componency relations with one another (Craver 2015; Craver and Bechtel 2007) (I briefly return to this in Section 2.4).

Mechanistic levels are necessarily local, in contrast to the more monolithic levels of Oppenheim and Putnam (1958), who divided nature into levels of atoms, molecules, cells, organs, organisms, and societies. For mechanistic levels, an entity is at a lower mechanistic level than another if and only if it is a component in the mechanism of the latter. From this a weak notion of sameness of level is derived: two entities are at the same level only if they are components in the same mechanism, and neither is a component of the other.[7]

[7] Eronen (2015) argues that this is so weak that it is tantamount to abandoning the idea of levels altogether.

A component of a mechanism is more than just a mereological part; it is a part that contributes to the behavior of the mechanism – it is a constitutively *relevant* part. There is some debate over how to cash out this notion of constitutive relevance. Craver (2007) characterizes it in terms of mutual manipulability of part and whole: A part is a component of (or is constitutively relevant to) a mechanism's behavior if one can manipulate that behavior by manipulating the part, and one can manipulate the part by manipulating the behavior of the mechanism. This account is not without problems[8], but I will not examine those here. Instead, I will assume that the notion of constitutive relevance as contribution to the behavior of a mechanism is clear enough for our purposes.

2.2.2 Mechanistic Explanation

The contemporary account of mechanistic explanation has its origin primarily in the work of Salmon (1984, 1989), among others (see, for example, Scriven [1959, 1975]). He developed a causal account of explanation in response to problems that arose for the deductive-nomological account (DN; also known as the covering-law model). According to DN, an explanation is an argument with descriptions of at least one law of nature and antecedent conditions as premises and a description of the explanandum phenomenon as the conclusion (Hempel and Oppenheim 1948). To explain is to show that the explanandum phenomenon is *predictable* given the truth of the premises. However, tying explanation this closely to prediction generates some now-infamous problems (Salmon 1989). For example, on such an account, many mere correlations come out as explanatory, which intuitively is not true. A falling barometer reliably predicts the weather, but the falling barometer does not explain the weather (Salmon 1989, 47).

[8] For example, Craver's account requires that the interventions used to establish constitutive relevance are 'ideal,' which seems conceptually impossible (Baumgartner and Gebharter 2015; Couch 2011; Harinen 2014; Leuridan 2012; Romero 2015).

According to Salmon's (1984) causal-mechanical view, in contrast, explanation involves situating the explanandum phenomenon in the causal structure of the world. (Salmon called this an 'ontic conception' of scientific explanation, contrasting it with the 'epistemic conception' of the deductive-nomological account. I return to this still-relevant distinction in Section 1.5.) There are several ways of so situating an explanandum phenomenon. An etiological causal explanation is "backward-looking;" it describes the explanandum phenomenon's past causal history (its immediately prior causes, the causes of those causes, and so on). A constitutive mechanistic explanation is "downward-looking;" it involves describing the entities, activities, and organization of the mechanism that produces, underlies, or maintains the explanandum phenomenon. It is the kind of explanation that most readily comes to mind when one hears the phrase "mechanistic explanation." However, there is also the neglected contextual mechanistic explanation, which is "upward-looking" (though see Craver [2001] and Bechtel [2011], from which I have borrowed the "looking" metaphor). It describes the broader mechanism of which the explanandum phenomenon is an active component[9]. This chapter will be concerned with the latter two kinds of explanation, constitutive and contextual mechanistic explanation (though I briefly return to etiological causal explanation in Section 2.3).

There are two important normative distinctions (or continua) in the mechanist's conceptual framework: mechanism schemata versus mechanism sketches, and how-possibly versus how-actually models (Machamer, Darden, and Craver 2000; Craver 2007). A mechanism schema is an abstract description of a *type* of mechanism, rather than a specific token instance. Details will inevitably be omitted, but, ideally, only details that are irrelevant to the mechanism

[9] Bechtel (2011) also includes what he calls "looking around," which involves determining how the components of a mechanism are organized. I have subsumed this under constitutive mechanistic explanation.

type. Details that are specific to tokens of the type can be added as the schema is applied to instances (Machamer et al. 2000, 15). Mechanism sketches, on the other hand, are incomplete descriptions of (type or token) mechanisms that contain black boxes and filler terms (Craver 2007, 113). They are still partially explanatory, but they are lacking in relevant detail. More details can be added to the model to fill in the gaps, though no model is ever fully complete, just complete enough for practical purposes (Craver and Darden 2013). Idealized models qualify as mechanism schemata, rather than sketches, to the extent that they capture relevant aspects of mechanisms.

A how-possibly model describes a merely possible mechanism, whereas a how-actually model describes the mechanism that (we have the most evidence to believe) actually produces, maintains, or underlies the explanandum phenomenon. This distinction is epistemic: turning a how-possibly model into a how-actually model does not require modifying the model itself in any way; it requires testing the model (Weiskopf 2011). The greater the evidential support for the model, the more how-actually it is. Between how-possibly and how-actually is a range of how-plausibly models. Turning a mechanism sketch into a more complete mechanism schema, in contrast, requires modifying the model by filling in missing details (Craver and Darded 2013). These details may be at the same mechanistic level as the rest of the details in the model, or they may be at a lower mechanistic level.

In contrast to how-possibly and how-actually models, or mechanism sketches and schemata, which more or less completely describe possible or actual mechanisms responsible for some explanandum phenomenon, a merely descriptive, or phenomenal, model describes an explanandum phenomenon, usually in a general, concise way. Snell's law is a common example

of a phenomenal model (Craver and Darden 2013). It accurately and compactly describes the relationship between the angle of incidence and the angle of refraction when light passes between two media, but it does not explain refraction.

Mechanistic explanations satisfy what are widely considered, by mechanists and non-mechanists (e.g., Chirimuuta 2014; Rice 2015; Weiskopf 2011) alike, the normative constraints on explanation: the ability to answer counterfactual questions about the explanandum phenomenon ('what-if-things-had-been-different' questions or, more compactly, w-questions), and the ability to manipulate and control[10] the explanandum phenomenon (Craver 2007). These norms capture what is distinctive about the scientific achievement of explanation, as opposed to other achievements like prediction, description, or categorization. As the barometer example above shows, a model can be predictive without being explanatory. They also provide a basis for explanatory power. A model is more explanatorily powerful, according to mechanists, when and only when it can answer more w-questions and afford more opportunities for control (Ylikoski and Kuorikoski 2010).

2.3. Models, Strategies, and Explanations

I have briefly described what mechanisms and mechanistic explanations are, but I have not yet given any examples of models that are mechanistic explanations. Canonical examples of mechanistic explanation have given the impression that a mechanistic explanation should look a certain way, or be constructed using certain methods, but some mechanists deny this (Craver 2014; Piccinini and Craver 2011; Zednik 2011; 2015).

In some of the most seminal work on mechanistic explanation (e.g., Bechtel and Richardson 1993; Glennan 1996; Machamer et al. 2000), the examples and diagrams used were

[10] These are related, of course. The latter ability is a practical analogue of the former.

very machine-like: biological oxidation, voltage-gated ion channels, the action potential, protein synthesis. This arguably led to the impression that a mechanistic explanation was a particular, machine-like kind of model or representation (Hochstein 2016; Zednik 2015).[11] With this impression in place, counterexamples from psychology (and elsewhere) have come in the form of explanatory models that look nothing like the mechanists' canonical examples. Implicit or explicit in many mechanists' responses to these counterexamples is a distinction between mechanistic explanations and mechanistic models.[12] Let us examine in some detail the dialectic in one prominent case from psychology – dynamical systems models – with that distinction in mind.

2.3.1 Dynamical Systems Models

Dynamical systems models are models that employ the mathematical (and geometric) concepts of dynamical systems theory, such as differential or difference equations (Chemero 2009; Izhikevich 2007; Zednik 2011). This allows the modeling of the time evolution of relevant variables, which can be represented geometrically (and graphically) as a trajectory through a phase or state space. The state space of a system represents all its possible states (i.e. all possible values of the system's variables). A trajectory through state space is then a graphical representation of how the system's variables change over time. Graphical representations have

[11] Although not all representations are models, in this context I will use the terms "representation" and "model" synonymously to mean some kind of structure (e.g., a concrete replica, a mathematical equation, a diagram, or a linguistic description) that is interpreted to represent a target system (Weisberg 2013). This terminological choice runs roughshod over Weisberg's distinction between models and model descriptions, but this should not affect the points that follow.

[12] I do not mean to imply that all mechanists make this distinction or that no mechanist has meant a certain kind of model by "mechanistic explanation," just that, of those mechanists who dispute putative examples of non-mechanistic explanation, many of them have appealed to this distinction, or something like it. Kaplan's (2011) model-to-mechanism-mapping (3M) requirement might preclude him from making this distinction, or at least from saying that models that fail 3M, but still provide information about mechanisms, are mechanistic explanations. For example, you can provide information about mechanisms by saying what is *not* responsible for an explanandum (cf. Lewis 1986: 220), but this would violate 3M. This comes down to Kaplan's intended scope of 3M, something about which I will refrain from speculating here.

the benefit of allowing careful and intuitive analysis of state space topology, revealing abstract, dynamical features such as the presence of attractors (i.e. states into which the system tends from surrounding states) (Izhikevich 2007). In dynamical models in psychology, the relevant variables often span brain, body, and environment (van Gelder 1998; van Gelder and Port 1995; Zednik 2011). I briefly describe two dynamical models: the HKB model and Beer's model of categorical perception.[13]

One of the first dynamical models that was presented as a challenge to mechanistic explanation was the Haken-Kelso-Bunz (HKB; Haken, Kelso, and Bunz 1985) model (Chemero 2009; Chemero and Silberstein 2011; Stepp, Chemero, and Turvey 2011; Walmsley 2008). Although the explanandum of this model is not an especially cognitive phenomenon, it will be helpful to review it and mechanists' responses.

HKB is a model of bimanual coordination, specifically simultaneous, side-to-side movement of the index fingers (and hands). The behavioral data were obtained by asking participants to move horizontally both index fingers either in-phase (pointing toward the midline, then away) or out-of-phase (both pointing left, then both right). Participants were asked to keep pace with a metronome so that experimenters could manipulate the rate of finger movement (Kelso 1981). By increasing the rate, experimenters found that only in-phase movement is possible beyond a certain critical rate. Participants who began out-of-phase involuntarily switched to in-phase once the critical rate was crossed. The same phenomenon occurs during other forms of bimanual coordination, such as hand movements at the wrist (Kelso 1984).

To model this phenomenon with dynamical systems theory, the fingers are represented as

[13] Since my theme is psychology, I leave aside dynamical models in neuroscience, though they too have been presented as counterexamples to mechanistic explanation. For example, see Ross (2015), which relies on Batterman and Rice's (2014) notion of a "minimal model explanation." The response to Batterman and Rice in Povich (forthcoming a) applies to Ross's argument as well.

coupled oscillators and the stable in-phase and out-of-phase movements as attractors. The dynamics are described by the following differential equation:

$$d\phi/dt = -dV/d\phi = -a \sin \phi - 2b \sin 2\phi,$$

where V is the so-called potential function, $V(\phi) = -a \cos \phi - b \cos 2\phi$, and the ratio b/a is a control parameter that varies inversely with finger oscillation frequency, and determines the topology of the phase space (i.e. the landscape of attractors). At a low oscillation frequency, there are two attractors, corresponding to stable in-phase and out-of-phase movement. At a high frequency, past the critical value, the landscape shifts to include only one attractor, corresponding to stable in-phase movement. This accurately describes the observed behavioral data.

Beer's (1996; 2003) dynamical model of perceptual categorization (or categorical perception) is more cognitively interesting (Zednik [2011] provides a detailed analysis of this model). The model is a simulated system consisting of a 14-neuron continuous-time recurrent neural network (CTRNN) brain, inside an evolved model agent (meaning its network architecture was constructed with an evolutionary algorithm[14]), inside a two-dimensional environment. The agent moves horizontally as circles or diamonds fall from above. It 'categorizes' these objects by catching the former and avoiding the latter. The agent perceives with an eye consisting of seven rays, each connected to a corresponding sensory input neuron. When a ray hits an object, its input neuron receives a signal inversely proportional to the distance from the object – the closer the object when 'seen' by a ray, the greater its input signal.

The agent with the best performance evolved a strategy of active scanning (Beer 2003). First, the agent centers the object in its field of view, then it moves back and forth, scanning the

[14] Specifically, the connection weights, biases, time constants, and gain were evolved, but not the number of nodes (Beer 2003, 214).

object. The scan narrows to hone in on circles, while breaking to avoid diamonds. Beer (2003, 228–9) explains this active scanning as follows. First, he decomposes the agent-environment dynamics into the effect of the relative positions of agent and object on the agent's motion, and vice versa. Then, for both circle and diamond trials, he superimposes the motion trajectory of the object through the agent's field of view onto a steady-state velocity field, which represents, for each point in the agent's field of view, the agent's steady-state horizontal velocity in response to an object at that point (228). Finally, he notices from an examination of the agent's motion trajectories that it consistently overshoots the midline of its visual field, due to the lag in time for the neural network to respond to sudden changes in sensory input. Therefore, according to Beer, active scanning is explained by the dynamic interaction of the steady-state velocity fields and the neural network's lag.[15]

Dynamicists have argued that dynamical models such as the above are non-mechanistic because they abstract from low-level neural details and capture high-level qualitative behavior, yet still explanatory because they yield accurate predictions and accurately describe, thereby unifying, diverse systems (Chemero 2009; Chemero and Silberstein 2011; Stepp, Chemero, and Turvey 2011; van Gelder and Port 1995; Walmsley 2008). HKB, for example, does not include any specification of the neural mechanisms responsible for finger movements, but it does accurately describe diverse systems (including the coordinated limb movements of two separate people [Schmidt, Carello and Turvey 1990]) and accurately predicts the amount of time it takes for the relative phase to stabilize following selective interference (Walmsley 2008).

2.3.2 Mechanist Responses

[15] Beer (2003) also analyzes the neural network, including individual neurons, and how it changes over the course of active scanning, thus providing a *multilevel* mechanistic explanation of categorical perception via active scanning. For brevity's sake, and due to the fact that this part of Beer's analysis is more clearly consistent with mechanistic explanation, I omit further discussion of this; see Zednik (2011) for more.

The responses of mechanists to dynamical models have invoked the distinction laid out in Section 1.2.2 between predictive, phenomenal models and mechanism-schemata. Although a dynamical model's predictive power is a virtue, it is not enough to make it explanatory (as the barometer example above shows). Similarly, Kaplan and Craver (2011) argue that a dynamical model's ability to apply to a wide range of diverse systems is insufficient for explanation. Instead, a model like HKB, insofar as any internal causal structure is omitted[16], is a phenomenal model that merely describes an interesting, widespread pattern, but does not explain that pattern. In light of these concerns, Kaplan and Craver (2011) argue that dynamicists have not yet provided a satisfactory account of what makes dynamical models explanatory, if they do not refer in any way to mechanisms or their organization (see also Kaplan 2015; Kaplan and Bechtel 2011).

Beer (2003) seems not to have explicitly taken his dynamical explanation to be non-mechanistic (see fn. 15). As Zednik (2011) argues, Beer's explanation should be seen as describing interactive components in a mechanism that spans brain, body, and environment. The explanandum is the behavior of one component in this mechanism, the agent's active scanning. The model shows how interactions with the environment, along with the time lag in responding to stimuli, result in active scanning.[17] While this explanation does not describe any internal mechanisms, so is not a constitutive mechanistic explanation, it does qualify as a contextual mechanistic explanation. Therefore, there appear to be some dynamical models that are also mechanistic explanations.[18]

---

[16] Kaplan and Craver (2011) note that Kelso and colleagues have not neglected to investigate the neural mechanisms that generate the dynamics HKB describes.

[17] Questions like, “Why is the lag such and such amount of time?” require looking at the neural mechanisms of the agent. This does not detract from the contextual mechanistic explanation of active scanning – the lag time is simply a different explanandum. See Zednik (2011: 254).

[18] Chemero (2009: xi, 85) argues that Gibson's ecological psychology (Gibson 1979) provides a background theory

Zednik (2011) makes an increasingly common distinction between mechanistic explanations, on the one hand, and the tools used for constructing and representing them, on the other. He reiterates that dynamical systems theory is a mathematical and conceptual framework that, as such, can be used to represent anything to which its concepts apply. If that includes the components, activities, and organization of mechanisms, then dynamical systems theory can provide mechanistic explanations.[19] Zednik (2015) has since extended this point, using examples from evolutionary robotics and network science to show how new tools for mechanism description and discovery go beyond the traditional strategies of decomposition and localization (Bechtel and Richardson 1993).

2.3.3 More on Models and Strategies

Hochstein (2016) hits on a distinction similar to Zednik's (2015) in his diagnosis of the disagreement over which models are mechanistic explanations. He locates two opposing assumptions concerning the role of representation in mechanistic explanation. He calls these assumptions the "representation-of" and "representation-as" accounts of mechanistic explanation. According to the representation-of account, for an explanation to be mechanistic, it must be a representation of a mechanism, where this requires only the provision of information about a mechanism. According to the representation-as account, for a explanation to be mechanistic, it must not only provide information about a mechanism, but also represent the mechanism mechanistically, that is, as a mechanism. That is, not only must the represented thing in the world be a mechanism, it must be represented in a particular way, mechanistically; the

unifying all dynamical modeling in psychology. Bechtel (2011) argues that Gibson's ecological psychology provides contextual mechanistic explanations.

[19] Similar arguments have been made with respect to network and graph theory: they can be used to provide mechanistic explanations when they capture the organizational features of a mechanism that are relevant to an explanandum (Levy and Bechtel 2013; Craver forthcoming).

model or representation itself must have the particular form of depictions of neatly localized entities interacting to produce the explanandum.[20]

On the representation-of account, the general relation between mechanistic explanations and models is as follows. In the world there is a target mechanism, that produces, maintains, or underlies an explanandum phenomenon. There are many, conceptually distinct ways of describing this mechanism. To the extent that a model accurately picks out the ontic structures relevant to the explanandum phenomenon, the model explains the explanandum phenomenon, regardless of how it is represented (and regardless of which concepts are deployed, if the representation is linguistic). The form of representation (and concepts deployed) become much more important when we are concerned with the understanding it provides to cognitive agents. Explanation and understanding should be kept relatively distinct; the concepts deployed in an explanatory text are more important for the latter than the former.

The representation-of account places no requirements on the form of the representation. Since the representation-as account requires more of a model for it to be a mechanistic explanation, fewer psychological models will be counted as mechanistic explanations according to it than according to the representation-of account. Here we see, then, how the two opposing assumptions lead to disagreement about which psychological models are mechanistic and why.

Hochstein (2016) argues that the representation-of account trivializes the claim that neuroscience provides mechanistic explanations. Since the brain is a collection of mechanisms and neuroscientists model the brain, they therefore provide mechanistic explanations. However, showing how neuroscientists provide mechanistic explanations requires showing how their

[20] One could also put this by saying there is a distinction between mechanistic models and models of mechanisms (Craver, personal commmunication).

concepts provide information about mechanisms, which is a controversial and nontrivial philosophical task, especially for computational and systems neuroscience (Piccinini and Craver 2011; Kaplan 2011; Povich 2015; Zednik 2015; I return to this in the next section). The same is true of etiological causal explanation. For example, Skow (2014) holds an account of causal explanation somewhat analogous to the representation-of account: roughly, an explanation is causal if and only if it provides information about the explanandum's causal history.[21] Skow responds to some prominent putative counterexamples to causal explanation (e.g., explanations that cite causally inert entities) by showing in detail how they provide causal information (e.g., how such explanations rule out possible causal histories).

The representation-of account is therefore not without some precedent. An account like Skow's (2014) has long been widely recognized as legitimate in the literature on causal explanation, where, to be a causal explanation, a representation need only provide information about an explanandum's causes or causal history, not identify any of its actual causes (Jackson and Pettit 1990; Lange 2013; Lewis 1986; Skow 2014). Furthermore, causal explanations do not have to have a particular form. Proponents of the representation-of account can be seen as extending this idea to mechanistic explanation.

A somewhat similar distinction is made by Levy (2013), who distinguishes between what he calls "causal mechanism," "explanatory mechanism," and "strategic mechanism." Only the latter two concern us here. According to Levy, explanatory mechanism is the thesis that "to explain a phenomenon, one must cite mechanistic information" (100). This appears to be what Hochstein (2016) would call a representation-of account. On the other hand, strategic mechanism "articulates a way of doing science, a framework for representing and reasoning about complex

[21] Skow's (2014) account is more complicated than this, and it is limited to explanations of particular events.

systems," using modeling methods such as decomposition and localization (104–5). Unlike the representation-as account, strategic mechanism does not explicitly say that mechanistic explanations must have a certain representational form, but such strategies do constrain the representational form of models. Adherence to strategic mechanism might therefore lead to adherence to the representation-as account.

2.4. Abstraction, Functionalism, and Realization

In addition to dynamical models, functional analyses are prominent putative counterexamples to mechanistic explanation in psychology (Fodor 1965; 1968; see Piccinini and Craver [2011] for response). A functional analysis of a psychological capacity explains it in terms of the functional properties, either of the whole cognitive system, or of its parts. Functional analysis is thought, by non-mechanists, to proceed relatively independently of consideration of the structural components that realize the functional properties, or play the functional roles. Mechanists have argued that functional analyses are mechanism-sketches (Piccinini and Craver 2011; Piccinini 2015; Povich 2015; call this the "sketch thesis"), while functionalists deny this (Weiskopf 2011; forthcoming). Let us examine more closely the reasons for and against the mechanistic status of functional analyses, which will bring out how realization and abstraction relate to mechanistic explanation.

2.4.1 Functional Analyses and Mechanism Sketches

The primary reason that Piccinini and Craver (2011) give for the sketch thesis is that functional analyses put constraints on the possible mechanisms that implement the functions identified in the analysis. Similarly, structure constrains function: not just any structural component can perform any function. For example, to perform the functions of belief and desire

boxes, a mechanism(s) must be able to distinguish between those two types of representation and transform them in relevant ways (Piccinini and Craver 2011, 303). This puts some constraints on what could possibly implement belief and desire boxes. This argument appears to rely on a representation-of account of mechanistic explanation (Hochstein 2016): functional analyses are mechanism sketches because they provide some information about mechanisms.

The neural mechanisms that play the functional roles of belief and desire boxes (or attentional filters or whatever), are likely vague, widely distributed, and multi-functional. For this reason, Piccinini and Craver (2011) also emphasize a permissive concept of mechanism like the one given in Section 2.2.1, according to which the components that play the functional roles need not be neatly localizable or contained within well defined boundaries (Piccinini and Craver 2011, 296).

Weiskopf (2011; forthcoming) objects to the sketch thesis and the claim, required for that argument, that mechanism components can be widely distributed. Against the latter claim, he argues that it results in "greater difficulty in locating the boundaries of mechanisms" (Weiskopf 2011, 315) and gives up "any requirement that parts be describable in a way that our modeling techniques can capture" (forthcoming, 22). I do not have space to respond in detail here, but I have argued in depth elsewhere that model-based fMRI can ameliorate these worries (Povich 2015).[22]

In response to the sketch thesis, Weiskopf (forthcoming) argues as follows. If functional analyses are mechanism-sketches, then they are amenable to two kinds of elaboration (24–5). Intralevel elaboration involves adding details, discharging filler terms, and so on, while staying

[22] Model-based fMRI is a neuroimaging method that combines psychological models with fMRI data, allowing cognitive neuroscientists to explore how the components of psychological models map onto distributed brain regions.

at the same mechanistic level. Interlevel elaboration involves going down mechanistic levels in order to explain their component entities and activities. He argues that it cannot be the case that functional analyses need interlevel elaboration in order to completely describe the causal structure relevant to a psychological phenomenon, because that would lead to a downward regress. In order to provide a complete model at any mechanistic level, one would have to give a complete model at every lower mechanistic level (26–7). He argues that if functional analyses need intralevel elaboration, this can be accomplished with more specific functional analyses of subsystems – there is no reason to think functional concepts can never fully accurately capture the psychological properties of a system (25).

The mechanist can make several moves in response to this argument. First, it could be argued that even if a functional analysis fully captures that psychological properties of a system, its explanatory incompleteness is shown by the fact that adding implementation details increases the explanatory power of the model (i.e. its ability to answer w-questions and afford opportunities of intervention and control). Adding implementation details need not always be a kind of interlevel elaboration either: to simply identify the occupant of a functional role is not to explain how that occupant plays its role. Endicott (2011) helpfully distinguishes between the "what" and the "how" of functional realization; only the latter requires descending mechanistic levels.

Second, the mechanist could accept that functional analyses can be complete mechanism schemata, rather than mere sketches. This seems a natural move for a proponent of the representation-of account, according to which the concepts deployed in an explanation do not affect its mechanistic status. The mechanist could argue that as long as the functional concepts

pick out features of mechanisms, functional analyses count as mechanistic explanations. This is just to deny the representation-as account that is presupposed in Weiskopf's argument, for example, when he claims that, “The question is whether remedying this sketchiness requires stepping out of the explanatory framework of psychology” (24) or that, “An ideally complete cognitive model will still be one that is couched in the autonomous theoretical vocabulary of psychology” (26). The argument that functional analyses are mechanism-sketches was not meant to imply that functional or computational analyses are never true or explanatory (Piccinini and Craver 2011). The argument was that functional analyses are true to the extent that they accurately describe mechanisms, and it is in virtue of being accurate descriptions of mechanisms that they explain. A central part of that argument was showing that different kinds of functional analysis are “elliptical” descriptions of mechanisms. However, if this is right, then it seems that Piccinini and Craver (2011; Piccinini 2015) were wrong that all functional analyses are mechanism-sketches. They seem to have the conceptual resources to say that sometimes functional analyses can be complete(-enough) mechanism schemata. Here again the key disagreement seems to be over the representation-of/representation-as account.

2.5. Ontic and Epistemic Conceptions of Explanation

The previous distinctions set out in this chapter are related to another that is prominent in contemporary philosophy of explanation: Salmon's (1984; 1989) distinction between epistemic, modal, and ontic conceptions of scientific explanation. These conceptions were different accounts of what a scientific explanation aims to show of its explanandum phenomenon: that it is expected to occur, that it had to occur, that it fits “into a discernible pattern” (1984, 121). According to Salmon, the “discernible pattern” into which an explanandum phenomenon is fit is

structured by causal processes, interactions, and laws (1984, 132). Explaining is "providing information about these patterns that reveals how the explanandum-events fit in" (1989, 121). Explanation, for Salmon, is not about nomic expectability or nomic necessity, but about fitting the explanandum into "discernible patterns" and "relationships that exist in the world" (1984, 121) (Povich forthcoming a).[23]

The ontic-epistemic debate has shifted twice since Salmon (Illari 2013)[24]. Salmon framed the debate in terms of what explanations do. After Salmon, the debate was framed metaphysically, as a debate about what explanations are: The ontic conception became the claim that scientific explanations are (almost always causal) dependence relations in the world; the epistemic conception became the claim that scientific explanations are epistemic states or representations (Povich forthcoming a).

The distinction has also shifted from a metaphysical distinction to one that focuses on explanatory demarcation and normative constraints on explanation (Craver 2014; Povich forthcoming a). Craver writes that according to the ontic conception, "in order to satisfy these two objectives [of explanatory demarcation and explanatory normativity], one must look beyond representational structures to the ontic structures in the world" (2014, 28). The idea is that attention to ontic structures, rather than representational form, is required to demarcate explanation from other scientific achievements, like prediction, and to distinguish good from bad explanations, how-possibly from how-actually explanations, and explanatorily relevant from irrelevant features (2014, 51). This formulation of the ontic conception has affinities with the representation-of account; in fact, it appears to just be the representation-of account.

[23] This need not be construed solely causally and Salmon did not think causation was essential to the ontic conception (See Chapter 4).

[24] The modal conception has fallen out of favor and was not included in later debates (but see Lange [2013] for a recent defense).

2.6. Conclusion

Which psychological models are mechanistic explanations? The conciliatory answer is, "It depends." It depends on whether one adopts a representation-of account or representation-as account (Hochstein 2016). I prefer making a distinction between mechanistic explanations and mechanistic models. Mechanistic explanations can be provided by non-mechanistic models, since non-mechanistic models can provide explanatory information about mechanisms. If you conceive the mechanistic project as "explanatory mechanists" (Levy 2013) tend to, as articulating a "downward" way of causally situating an explanandum phenomenon that was neglected by Salmon and others who focused on "backward" (etiological) causal explanation (Craver 2007, 8), then it becomes clearer why one might hold the representation-of account. A categorization of the diverse kinds of representation or model used in scientific explanatory practices might be useful for some purposes, but does not seem to advance the classical project of a philosophical theory of scientific explanation, which is to provide conditions of explanatory demarcation and explanatory normativity (Craver 2014).

Chapter 3. Mechanisms and Model-Based fMRI

Mechanistic explanations satisfy widely held norms of explanation: the ability to manipulate and answer counterfactual questions about the explanandum phenomenon. A currently debated issue is whether any non-mechanistic explanations can satisfy these explanatory norms. Weiskopf (2011) argues that the models of object recognition and categorization, JIM, SUSTAIN, and ALCOVE, are not mechanistic, yet satisfy these norms of explanation. In this paper, I argue that these models are mechanism-sketches. My argument applies recent research using model-based fMRI, a novel neuroimaging method whose significance for current debates on psychological models and mechanistic explanation has yet to be explored.

3.1. Introduction

A mechanistic explanation of a phenomenon describes the entities, activities, and organization of the mechanism that produces, underlies, or maintains that phenomenon (Bechtel and Abrahamsen 2005; Craver 2007). Mechanistic explanations satisfy what are widely considered the normative constraints on explanation: the ability to answer a range of counterfactual questions regarding the explanandum phenomenon and the ability to manipulate and control the explanandum phenomenon (Craver 2007). These norms capture what is distinctive about the scientific achievement of *explanation* as opposed to prediction, description, or categorization. A currently debated issue is whether any non-mechanistic forms of explanation can satisfy these explanatory norms.[25] Weiskopf (2011) argues that the models of object recognition and categorization, JIM, SUSTAIN, and ALCOVE, are not mechanistic explanations and nonetheless satisfy these normative constraints.

[25] Batterman and Rice (2014) is a provocative recent paper arguing affirmatively.

I argue that JIM, SUSTAIN, and ALCOVE are in fact, and are intended by their creators to be, mechanism-sketches, i.e. incomplete mechanistic explanations. My argument applies recent research using model-based fMRI. Model-based fMRI allows cognitive neuroscientists to locate even widely distributed neural components in psychological models. These novel neuroimaging methods have developed only recently (Glascher and O'Doherty 2010), and philosophers have yet to discuss their significance for current debates on psychological models and mechanistic explanation.

The paper is organized as follows. In Section 3.2, I motivate the mechanistic account of explanation and introduce two important distinctions in that account: complete models vs. mechanism-sketches, and how-possibly vs. how-actually models. In Section 3.3, I introduce the three models of object recognition and categorization that Weiskopf takes as the scientific grounds for his philosophical thesis. In Section 3.4, I present Weiskopf's arguments for thinking these models are non-mechanistic, yet explanatory. I also begin to respond to these arguments. I show precisely why JIM should be seen as a mechanism-sketch. In Section 3.5, I show how the inventors of SUSTAIN and ALCOVE have subsequently used model-based fMRI to decide between these mechanism-sketches on the basis of information about widely distributed parts.

3.2. Mechanistic Explanation

The mechanistic account of explanation developed out of Salmon's (1984) insight into the problems that arise when an account of explanation is tied too closely to prediction. Salmon's principal target was the deductive-nomological account. According to the deductive-nomological account (Hempel and Oppenheim 1948), an explanation is an argument with descriptions of at least one law of nature and antecedent conditions as premises and a description of the

explanandum phenomenon as the conclusion. On this view, to explain is to show that the explanandum phenomenon is predictable on the basis of at least one law of nature and certain specific antecedent and boundary conditions. However, tying explanation this closely to prediction generates some famous problems (Salmon 1989). On such a view, many mere correlations come out as explanatory. For example, a falling barometer reliably predicts the weather but the falling barometer does not *explain* the weather. In contrast, on the causal-mechanical view, explanation involves situating the explanandum phenomenon in the causal structure of the world. There is more than one way of situating a phenomenon in the causal structure of the world, and in this paper I am solely concerned with explanations that identify the mechanism that produces, underlies, or maintains the explanandum phenomenon.[26]

If one ties explanation so closely to prediction, one risks missing what makes explanation a distinctive scientific achievement. Weiskopf (2011) and I in fact agree on what makes explanation distinctive: explanations provide the ability to answer a range of counterfactual questions regarding the explanandum phenomenon and the ability to manipulate and control the explanandum phenomenon. Weiskopf and I disagree about what kinds of explanation or model can satisfy these norms.

Within the mechanistic framework there are two important distinctions that will be necessary in the arguments that follow: complete models vs. mechanism-sketches, and how-possibly vs. how-actually models (Craver 2007). Mechanism-sketches are incomplete descriptions of mechanisms that contain black boxes and filler terms (Ibid., 113). They are still partially explanatory. More details can be added to the model to fill in the gaps, though no model

[26] See Bechtel (2009) for a discussion of some other ways of causally situating a phenomenon. What Bechtel calls "looking down," I am here calling "mechanistic explanation."

is ever fully complete, just complete enough for practical purposes. There can certainly be too many details for the purposes of the modeler and the details that *are* included should be relevant.[27] Idealized models can be mechanistic explanations even if they are in some sense incomplete; they can exclude irrelevant detail.

A how-possibly model describes a merely possible mechanism, whereas a how-actually model describes the mechanism that (we have the most evidence to believe) actually produces, maintains, or underlies the explanandum phenomenon. As Weiskopf (315) rightly points out, this distinction is epistemic. Turning a how-possibly model into a how-actually model does not require modifying the model itself in any way; it requires testing the model. The greater the evidential support for the model, the more how-actually it is. In contrast, turning a mechanism-sketch into a more complete mechanistic explanation requires modifying the model by filling in missing details.

3.3. JIM, SUSTAIN, and ALCOVE

In this section, I introduce the models of object recognition and categorization on which Weiskopf builds his case for the existence of non-mechanistic yet explanatory models. In Section 3.4, I present Weiskopf's arguments for thinking these models are non-mechanistic, yet explanatory.

According to JIM (John and Irv's Model), in perception objects are broken down into viewpoint-invariant primitives called "geons". Geons are simple three-dimensional shapes such as cones, bricks, and cylinders. The properties of geons are intended to be non-accidental properties (NAPs), largely unaffected by rotation in depth (Biederman 2000). Objects are represented as spatially arranged collections of geons. The geon structure of perceived objects is

[27] See Craver (2007, 139-60) for one account of constitutive (i.e. mechanistic) relevance.

extracted and stored in memory for later use in comparison and classification.

The importance of NAPs is shown by the fact that sequential matching tasks are extremely easy when stimuli differ only in NAPs. If you are first shown a stimulus, then a series of rotated stimuli, each of which differs from the first only in NAPs, it is a simple matter to judge which stimuli are the same as or different from the first. Sequential matching tasks with objects that differ in properties that *are* affected by rotation in depth are much harder.

In JIM, this object recognition and categorization process is modeled by a seven layer neural network (Biederman, Cooper, and Fiser 1993). Layer 1 extracts image edges from an input of a line drawing that represents the orientation and depth of an object (182). Layer 2 has three components that represent vertices, axes, and blobs. Layer 3 represents geon attributes such as size, orientation, and aspect ratio. Layers 4 and 5 both derive invariant relations from the extracted geon attributes. Layer 6 receives inputs from layers 3 and 5 and assembles geon features, e.g., "slightly elongated, vertical cone above, perpendicular to and smaller than something" (184). Layer 7 integrates successive outputs from layer 6 and produces an object judgment.

ALCOVE (Attention Learning Covering map), like JIM, is a neural network model of object categorization (Kruschke 1992). It has 3 layers. The perceived stimulus is represented as a point in a multidimensional psychological space with each input node representing a single, continuous psychological dimension. For example, a node may represent perceived size, in which case the greater the perceived size of the stimulus, the greater the activation of that node. Each node is modulated by an attentional gate whose strength reflects the relevance of that dimension for the categorization task. Each hidden node represents an exemplar and is activated

in proportion to the psychological similarity of the input stimulus to the exemplar. Output nodes represent category responses and are activated by summing hidden nodes and multiplying by the corresponding weights.

SUSTAIN (Supervised and Unsupervised Stratified Adaptive Incremental Network) is a neural network model similar to ALCOVE (Love, Medin, and Gureckis 2004). Its input nodes also represent a multidimensional psychological space, but they can take continuous and discrete values. Like ALCOVE, inputs are modulated by an attentional gate. Unlike ALCOVE, which stores all items individually in memory in exemplar nodes, the next layer of SUSTAIN consists of a set of clusters (bundles of features) associated with a category. Each cluster activates in proportion to its proximity to the input in multidimensional psychological space; the more similar a cluster is to the input, the more it activates. There are inhibitory connections between each cluster, so that the cluster most similar to the input inhibits all others. This winning cluster activates the output unit generating the category label.

3.4. Weiskopf's Arguments

Weiskopf argues that the previous models are able to satisfy the norms of explanation but are not mechanistic models. How do these models provide the ability to answer counterfactual questions about the explanandum phenomenon, and the ability to manipulate and control the explanandum phenomenon? According to Weiskopf, they satisfy these explanatory norms "because these models depict one aspect of the causal structure of the system" (334). This claim is *prima facie* in tension with Weiskopf's claim that these models are not mechanistic. He argues, "[T]here may be an underlying mechanistic neural system, but this mechanistic structure is not what cognitive models capture." (333).

One way of reconciling the above claims is to argue that these models are explanatory because they depict causal structure, but they are not mechanistic because the causal structure that they depict is not a mechanism. This is the line Weiskopf takes. Why, according to Weiskopf, are these causal structures not mechanisms? He argues,

> If parts [of mechanisms] are allowed to be smeared-out processes or distributed system-level properties, the spatial organization of mechanisms becomes much more difficult to discern. … Weakening the spatial organization constraint by allowing distributed, nonlocalized parts incurs costs, in the form of greater difficulty in locating the boundaries of mechanisms and stating their individuation conditions. (334)

The causal structures depicted by JIM, SUSTAIN, and ALCOVE should not be thought of as mechanisms, according to Weiskopf, because the structures that putatively implement them are highly distributed. If mechanisms are allowed to contain distributed, non-localized parts, this will make it difficult to locate them. Call this the practical problem of non-localization. Weiskopf does not provide any reason to think that the *philosophical* (rather than practical) problem of mechanism individuation is made more difficult by allowing distributed parts or that existing accounts[28] of mechanism individuation cannot handle distributed parts.[29] Yet numerous neuroimaging methods, especially model-based fMRI, ameliorate this practical problem. Model-based fMRI is well-suited to mechanistically discriminate between competing, equally behaviorally confirmed[30] psychological models.

In addition to Weiskopf's practical problem, there is what I will call the triviality problem

---

[28] See fn. 27 for an account of mechanism individuation.

[29] Weiskopf (331) also cites the phenomenon of neural reuse as inconsistent with mechanistic explanation, but the fact that a part of one mechanism can also be a part of a different mechanism constitutes only a practical problem for mechanism individuation.

[30] Weiskopf (335–6) is right that evidence for psychological models can come from many places. Although psychological models can be supported and constrained behaviorally, this degree of "evidential autonomy" does not establish the *explanatory* autonomy Weiskopf requires. It does not affect the mechanist's point that the parts of a psychological model must correspond to brain regions that implement the relevant computations for the model to be explanatory.

of non-localization. Weiskopf argues that if these kinds of distributed part are allowed, then "it is far from clear what content the notion of a mechanism has anymore" (334). First, as I have said, there has been no argument that existing accounts of mechanism individuation cannot accommodate distributed parts. If these accounts are workable while allowing distributed parts, then the notion of a mechanism remains contentful. Second, this objection misunderstands the mechanistic project, or at least a plausible way of conceiving that project. If you conceive the mechanistic project as articulating a "downward" way of causally situating an explanandum phenomenon that was neglected by Salmon and others who focused on "backward" (etiological) causal explanation (Craver 2007, 8), then a "liberalization" of the notion of mechanism that permits distributed parts is perfectly in line with that project and should not be seen as any kind of concession or retreat. Although such a "liberalization" may make mechanisms even more ubiquitous than they already were, it does not make every physical system a mechanism. For example, mere aggregates lack the organization necessary to be mechanisms (Ibid., 135–39).

Next I will present some of the neuroimaging studies conducted with JIM and argue that JIM is a mechanism-sketch. JIM was built, not merely to produce the same behavior as human beings in object recognition tasks, but to model something that might really be happening in human brains (Biederman, Cooper, Hummel, and Fiser 1993, 176). Accordingly, Irving Biederman, one of the co-creators of JIM, and others have conducted various neuroimaging studies to investigate the neural underpinnings of the model.

If JIM is a mechanism-sketch, the systems and processes in the model required for the extraction, storage, and comparison of geon structures must to some extent correspond to (perhaps distributed) components in the brain's actual object recognition system. For example, if

JIM is a mechanism-sketch, there is an area or a configuration of areas in the brain where simple parts and non-accidental properties are represented. In one study investigating this (Hayworth and Biederman 2006), participants were shown line drawings that were either local feature deleted (LFD), in which every other vertex and line was deleted from each part, removing half the contour, or part deleted (PD) in which half of the parts were deleted. On each trial, participants saw either LFD or PD stimuli presented as a sequential pair and had to report whether the exemplar depicted by the second stimulus was the same as or different than that depicted by the first. The second stimulus was always mirror-reversed with respect to the first. Each experimental run was comprised of an equal number of three conditions: Identical, Complementary, and Different Exemplar. In the Identical condition, the second stimulus was identical to the first stimulus (though mirror-reversed). In the Complementary condition, the second stimulus depicted the same exemplar as the first, but the second stimulus was a "complement" of the first stimulus. An LFD-complement is composed of the deleted contour of the first stimulus and a PD-complement is composed of the deleted parts of the first stimulus. In the Different Exemplar condition, the second stimulus depicts a different exemplar than the first.

This study used an fMRI-adaptation design that relies on the assumption that when two successive stimuli activate the same brain region, neural activity reduces (Krekelberg, Boynton, van Wezel 2006, 250). The results of the study showed adaptation between LFD complements and lack of adaptation between PD complements in lateral occipital complex, especially the posterior fusiform area, an area known to be involved in object recognition. These results imply that this area is "representing the parts of an object, rather than local features, templates, or object concepts" (Hayworth and Biederman 2006, 4029). Biederman has conducted many other

fMRI experiments, including some that "suggest that LO [lateral occipital cortex] is the locus of the neural correlate for the greater detectability for nonaccidental relations" (Kim and Biederman 1824).

Though these experiments suggest that JIM should be seen as a mechanism-sketch, Weiskopf has another argument for why it should not: JIM has properties that do not and could not correspond to anything in the brain. Weiskopf (331) refers to JIM's "Fast Enabling Links" (FELs), which allow the model to bind representations and have infinite propagation speed. Weiskopf calls FELs an example of, "fictionalization," or, "putting components into a model that are known not to correspond to any element of the modeled system, but which serve an essential role in getting the model to operate correctly" (Ibid.). The FELs, Weiskopf argues, undermine the claim that JIM is a mechanism-sketch.

Weiskopf is right that FELs are an essential fictionalization. However, playing an essential role in getting a model to operate is not the same as explaining; these parts of the model carry no explanatory information and render the model, or at least part of it, how-possibly (where the possibility involved is not physical possibility, since FELs are physically impossible). FELs play the black box role of whatever-it-is-that-accounts-for-binding. In addition to playing a black box role, they serve practical and epistemic purposes such as suggesting, constraining, and sharpening questions about mechanisms (Bogen 2005). Let me explain how by comparing FELs to Bogen's example of the Goldman, Hodgkin, and Katz (GHK) equations.

The GHK voltage and current equations are used to determine the reversal potential across a cell's membrane and the current across the membrane carried by an ion. These equations rely on the incorrect assumptions that each ion channel is homogeneous and that interactions

among ions do not influence their flow rate (Bogen 409). Bogen highlights the effects on research of these incorrect assumptions,

> Investigators used these and other GHK equation failures as problems to be solved by finding out more about how ion channels work. Fine-grained descriptions of exceptions to the GHK equations and the conditions under which they occur sharpened the problems and provided hints about how to approach them. (410)

The GHK equations provide a case of "using incorrect generalizations to articulate and develop mechanistic explanations" (409). Something similar can be said about FELs. Not only do FELs play an essential black box role, FELs suggest new questions about mechanisms, new problems to be solved. For example, Hummel and Biederman (1992) write,

> [FELs allow] JIM to treat the constraints on feature linking (by synchrony) separately from the constraints on property inference (by excitation and inhibition). That is, cells can phase lock without influencing one another's level of activity and vice versa. Although it remains an open question whether a neuroanatomical analog of FELs will be found to exist, we suggest that the distinction between feature linking and property inference is likely to remain an important one. (510)

Like the GHK equations, FELs suggest new lines of investigation, in this case regarding the relation between feature linking, property inference, and their neural mechanisms. Specifically, FELs suggest research questions such as, "Can biological neurons phase lock without influencing one another's activity?" and "Are there other ways biological neurons could implement feature linking and property inference *independently*?"

In the next section, I will explain model-based fMRI and demonstrate how recent model-based fMRI research shows that, like JIM, SUSTAIN and ALCOVE are mechanism-sketches.

3.5. Model-Based fMRI

Functional magnetic resonance imaging (fMRI) is a neuroimaging method that provides an indirect measure of neuronal activity. More specifcally, fMRI measures a physiological

indicator of oxygen consumption that correlates with changes in neuronal activity (Huettel, Song, and McCarthy 159–160).

Model-based fMRI is a neuroimaging method that combines psychological models with fMRI data. It "provides insight into 'how' a particular cognitive function might be implemented in the brain, not only 'where' it is implemented" (O' Doherty, Hampton, and Kim 39). In this way, model-based fMRI provides a way of discriminating between competing, equally behaviorally confirmed cognitive models (Glascher and O'Doherty 502). Furthermore, "the more complex the model (and hence the more associated free parameters), the more unconstrained the behavioral fitting becomes," in which case the additional constraints imposed by neurophysiological and neuroimaging data become "even more critical" (O' Doherty, Hampton, and Kim 37; White and Poldrack 2013).

To conduct a model-based fMRI analysis, one starts with a psychological model that postulates internal variables between stimulus input and behavioral output. While research participants perform a model-relevant task, researchers obtain fMRI data from which they can locate neural correlates of the internal variables (O' Doherty, Hampton, and Kim 36). The model-predicted values of internal variables across trials are convolved (mathematically combined) with a canonical hemodynamic response function (HRF) (Glascher and O'Doherty 505). This is done to account for the usual lag in the hemodynamic response (O' Doherty, Hampton, and Kim 37). This yields a new, model-predicted HRF that can be regressed against the obtained fMRI data. This allows researchers to identify brain areas where the model-predicted HRF significantly correlates with the observed HRF across trials.[31]

[31] Batterman and Rice (2014) object that the notion of *correspondence* between model and world is never explained by mechanists. I have no general theory of correspondence, but the sense in which (parts of) a psychological model correspond(s) to (parts of) the brain should be clear in each case. Here, for example, correspondence is significant correlation between model-predicted and observed HRF.

I should make clear that model-based fMRI inherits the limitations of fMRI, such as poor spatiotemporal resolution, and does not obviate the need for other neuroimaging methods (e.g., PET, EEG, or MEG), to which the model-based approach can also be applied.

Now that we have a basic understanding of how model-based fMRI works and what it can accomplish, let me return to SUSTAIN and ALCOVE and show how they are mechanism-sketches by drawing on recent model-based fMRI research.

Both models were investigated in a model-based fMRI study in which participants completed a rule-plus-exception category learning task (Davis, Love, and Preston 2012). During the task, a schematic beetle was presented and participants were asked to classify it as living in Hole A or Hole B. Participants then received feedback on the correctness of their classification. The beetles varied on four of the following five attributes, with the fifth held constant: eyes (green or red), tail (oval or triangular), legs (thin or thick), antennae (spindly or fuzzy), and fangs (pointy or round). Six of the eight beetles presented could be correctly categorized on the basis of a single attribute. For example, three out of four Hole A beetles had thick legs and three out of four Hole B beetles had thin legs. These were the rule-following beetles. The other beetles were exceptions to the rule, having legs that appeared to match the other category.

Two predictions from SUSTAIN and ALCOVE were tested. First, each model predicts specific changes in recognition strength across trials. During stimulus presentation SUSTAIN predicts a recognition advantage for exceptions; ALCOVE predicts no recognition advantage. This difference in recognition strength predictions arises because in ALCOVE, but not in SUSTAIN, all items are stored individually in memory regardless of whether they are exceptions or rule-following items. Second, each model predicts specific changes in error correction across

trials. The amount of error is given by the difference between the model's category response and the correct response. Both SUSTAIN and ALCOVE predict that exceptions will always produce more error than rule-following items, although both will produce less error as learning progresses (Ibid., 266).

The results revealed that both the recognition strength and error correction measures predicted by SUSTAIN found significant correlations in medial temporal lobe (MTL) regions including bilateral hippocampus, parahippocampal cortex, and perirhinal cortex. ALCOVE's predicted recognition strength measure did not find any significant correlations in MTL, although its predicted error correction measure found significant correlations in MTL regions (Ibid., 266–67). These results "suggest that, like SUSTAIN, the MTL contributes to category learning by forming specialized category representations appropriate for the learning context" (Ibid., 269).

SUSTAIN is more how-actually (evidentially supported) than ALCOVE because both of SUSTAIN's prediction measures (recognition strength and error correction) were significantly correlated with observed HRF, whereas only one of ALCOVE's prediction measures (error correction) was significantly correlated. These experiments also show that cognitive neuroscientists are currently advancing the ability to map the entities and activities in psychological models to distributed neural systems, such as MTL regions spanning bilateral hippocampus, parahippocampal cortex, and perirhinal cortex.

Davis, Love (a creator of SUSTAIN), and Preston (2012) are at times quite explicit that they are treating the models as mechanism-sketches. For instance, they write, "We use a model-based functional magnetic resonance imaging (fMRI) approach to test the proposed mapping

between MTL function and SUSTAIN's representational properties" (261). Given their emphasis on mapping models to the brain, it is clear that they intend these models to be mechanistic, as Biederman intends JIM to be. They are interested in more than the behavioral accuracy of these models; after all, SUSTAIN and ALCOVE are already behaviorally well-confirmed. The main difference between the two is in their hidden layers, where SUSTAIN has clusters and ALCOVE stores items individually. Model-based fMRI allowed Davis et al. to gather evidence relevant to assessing which of these models was more mechanistically accurate.

3.6. Conclusion

Weiskopf (2011) presents three models of object recognition and categorization, JIM, ALCOVE, and SUSTAIN, that he claims are non-mechanistic, yet explanatory. He argues that they are not mechanistic because their parts cannot be neatly localized and because they contain some components that cannot correspond to anything in the brain, such as Fast Enabling Links, but are nevertheless essential for the proper working of the model. I argue, on the contrary, that in addition to playing a black box role, FELs play useful, non-explanatory roles such as suggesting new lines of investigation regarding feature linking and property inference.

My argument for the claim that SUSTAIN and ALCOVE are mechanism-sketches relies partly on model-based fMRI research. Model-based fMRI and other model-based neuroimaging methods allow cognitive neuroscientists to explore how psychological models map onto the brain. This helps cognitive neuroscientists discriminate between equally behaviorally confirmed psychological models.

Biederman, Love, et al. treat JIM, SUSTAIN, and ALCOVE as mechanism-sketches, and they should. They should because by locating mechanisms one opens a new range of

opportunities for manipulating the mechanism and one obtains answers to counterfactual questions that were not available before. For example: What kinds of deficit in categorization performance would result from a lesion in bilateral hippocampus? If someone has a specific deficit in categorization performance, how might we fix it? Where might the problem lie? This increases the explanatory power of these models.

The development of these model-based approaches has broader implications, beyond the narrow dispute over JIM, SUSTAIN, and ALCOVE, for the debate over the explanatory and mechanistic status of psychological models. As cognitive neuroscientists continue to test competing models against neuroimaging data using model-based techniques, it is likely that they will, as they should, retain those models that are mechanistically accurate and discard those that are not, and in so doing reveal that explanatory progress in cognitive neuroscience consists in the development of increasingly mechanistic models.

Chapter 4. Minimal Models and the Generalized Ontic Conception of Scientific Explanation

Batterman and Rice (2014) argue that minimal models possess explanatory power that cannot be captured by what they call "common features" approaches to explanation. Minimal models are explanatory, according to Batterman and Rice, not in virtue of accurately representing relevant features, but in virtue of answering three questions that provide a "story about why large classes of features are irrelevant to the explanandum phenomenon" (356). In this chapter, I argue, first, that a method (the renormalization group) they propose that answers the three questions cannot answer them, at least by itself. Second, I argue that answers to the three questions are unnecessary to account for the explanatoriness of their minimal models. Finally, I argue that a common features account, what I call the generalized ontic conception of explanation, can capture the explanatoriness of minimal models.

4.1. Introduction

While acknowledging the widespread use of causal explanation in science, a number of prominent philosophers of science have recently begun exploring its limits (see Batterman 2002a; 2002b; Huneman 2010; Rice 2012; 2013; Woodward 2013). Batterman has been investigating the ways in which neglect of causes contributes to explanatory power in physics, particularly in statistical mechanics. Rice has been engaged in similar investigations of the neglect of causes in optimality modeling in biology. Recently, Batterman and Rice (2014; henceforth "B&R") have combined their efforts in an articulation of their common project. Their work brings important and successful modeling techniques to bear on the philosophy of scientific explanation. Nevertheless, there are significant limitations to their project. It is my aim here to

spell out these limitations and provide an alternative proposal.

B&R focus on minimal models, which are "used to explain patterns of macroscopic behavior across systems that are heterogeneous at smaller scales" (349). This widespread class of models, they argue, has explanatory power that cannot be captured by what they call "common features" approaches to explanation. According to common features approaches, 1) explanations accurately represent all and only[32] the features relevant to their explananda and 2) the explanatoriness of a representation consists in its representing relevant features (351).[33] Common features approaches include not only mechanistic approaches (Craver 2006; Glennan 2002; Kaplan 2011) and causal and difference-making approaches (Salmon 1984; 1989; Strevens 2008; Woodward 2003), but also Pincock's (2012) structuralist or mapping account, which explicates the explanatory role of mathematics in terms of its ability to mirror certain ontic structures. Any philosophical theory of explanation according to which accurate representation is responsible for explanatory power is a common features approach, whether or not the features represented are causes (B&R, 351).

B&R argue that common features approaches fail to capture the explanatoriness of minimal models because, even when a minimal model is minimally accurate, it is not its accuracy that accounts for its explanatoriness. Rather, minimal models are explanatory in virtue of "there being a story about why large classes of features are irrelevant to the explanandum phenomenon" (356).

---

[32] Depending on the explanatory representation used, some irrelevant features must be represented. For example, if our explanatory representation is pictorial, it must be colored some way, even if color is not relevant to the explanandum phenomenon. Ideally the modeler will flag any potential confusions. See Weisberg (2013, §3.3) for a related discussion of the role of modelers' intentions in determining what he calls 'representational fidelity criteria', standards for evaluating a model's representational accuracy.

[33] 1) is not just a restatement of 2). One could hold that accurate representation is necessary but not sufficient for explanation. This appears to be close to B&R's view (351, 356).

In this paper, I argue for a negative and a positive thesis. My negative thesis is that B&R's account of the explanatoriness of minimal models fails. They require that three questions be answered in order to provide the above-mentioned story about why large classes of features are irrelevant. I will henceforth refer to these as the 'Three Questions':

> Q1. Why are these common features necessary for the phenomenon to occur?
> Q2. Why are the remaining heterogeneous details (those left out of or misrepresented by the model) irrelevant for the occurrence of the phenomenon?
> Q3. Why do very different [fluids and populations] have features...in common?[34] (361)

My negative thesis consists of two parts. First, the method they propose to answer the Three Questions is unable to answer them, at least by itself. Second, answers to the Three Questions are unnecessary to account for the explanatoriness of minimal models. I argue for this second claim in two ways. First, I analogize their strategy to an exactly similar strategy in a more commonplace case of multiple realizability. In the case I present, it is evident that answering analogues of the Three Questions is unnecessary to explain multiple realizability. Second, I argue that if answers to the Three Questions were necessary, a regress would loom. B&R need to explain why, if the Three Questions are necessary, we should stop asking where they say we should. Of course, according to B&R, the Three Questions are not further questions, in addition to the question of what makes minimal models explanatory; the Three Questions just are those that need to be answered in order to account for the explanatoriness of minimal models. My analogy is intended to show that that is not the case.

My positive thesis is that a common features approach can account for the

[34] I have slightly altered the wording of Q3 to capture both models, thereby avoiding unnecessary repetition.

explanatoriness of minimal models.[35][36] B&R are (probably[37]) right that mechanistic and difference-making accounts cannot do the job, but an account much like the one proposed by Bokulich (2011), Rice himself (2013), and Saatsi and Pexton (2013) can. They follow Woodward (2003) in requiring that an explanation represent counterfactual dependence relations between the explanandum phenomenon and the features on which it depends, but they drop the requirement that these counterfactual dependence relations be construed causally. The reason for this is that the counterfactual dependence relations represented by some models, such as B&R's minimal models, cannot very plausibly be given a causal interpretation.

On this view, explanatory power consists in the ability to answer what-if-things-had-been-different questions ("w-questions"). I argue that this requires commitment to an ontic conception of scientific explanation (Salmon 1984) and that philosophers of science have been mistaken in equating the ontic conception with the causal-mechanical account of explanation. As we will see, Salmon seems not to have equated them.

My proposal is consistent with many things B&R have themselves written in the past.[38] It seems that their desire to avoid anything like a common features approach has driven them too far, apparently past things they have said before. In the present atmosphere in philosophy of science, it is a significant enough achievement to have brought to philosophical focus important modeling methods in physics and biology that emphasize the systematic neglect of causal detail.

[35] While I was finishing this manuscript, Lange (2015) also made this point, although he does not develop the positive proposal I do. He also made an objection to B&R similar to one of mine about regress. These and any other commonalities were arrived at independently.

[36] I also think that the common features that are shared between minimal models and real world systems are what justifies scientists' applications of the former to the later, though I do not have space to argue for this here.

[37] It is somewhat plausible that at least some of the common features in B&R's minimal models can be given a causal interpretation. On the account proposed here, though, this not what makes these features explanatory. I briefly expand on this at the end of Section 4.4.

[38] For examples, see fn. 56, Batterman's remarks on pain below, and Rice (2013): '*in some cases counterfactual information can be explanatory without tracking any relationships of causal dependence*' (20; original emphasis).

B&R have rightly stressed the importance of this neglect, but this importance need not drastically change our account of scientific explanation.

The rest of the paper is organized as follows. In Section 4.2, I present the minimal models whose explanatoriness B&R argue cannot be accounted for by a common features approach. These are the Lattice Gas Automaton (LGA) model of fluid dynamics and Fisher's model of 1:1 sex ratios. In Section 4.3, I present and critique B&R's account of the explanatoriness of these minimal models. According to B&R, any such account must answer the Three Questions, and answers are provided by the renormalization group (RG) and universality classes[39]. I argue that the Three Questions cannot in fact be answered by RG alone. I then argue that regardless of whether RG answers the Three Questions, they do not need to be answered in order to give an account of the explanatoriness of LGA and Fisher's model. I give two arguments for this. First, I show that answers to analogues of the Three Questions are unnecessary in an analogous case of multiple realizability. Batterman (2000) has argued that RG explains multiple realizability generally, so I take it that my analogy is apt and generalizable to B&R's models. Second, I argue that if answering the Three Questions were necessary for an account of the explanatoriness of B&R's minimal models, a regress would loom.

In Section 4.4, I provide my own common features account of the explanatoriness of B&R's minimal models: the generalized ontic conception. I argue that they are explanatory because they accurately represent the relevant dependence relations, that is, the objective features of the world on which the explanandum phenomenon counterfactually depends. My account is an ontic conception, in Craver's (2014) sense (to be explained more fully below). I argue, for

[39] Of course, in biological contexts some mathematical method(s) other than RG must be employed, though B&R are silent on what these methods might be.

reasons different than Wright (2012), that it is a mistake to equate the ontic conception of scientific explanation with the causal-mechanical account of explanation (Craver [2014] gestures at this idea in his defense of the ontic conception). A viable general theory of scientific explanation can be constructed by combining insights from Salmon (1984; 1989) and Woodward (2003), while realizing that there are noncausal kinds of ontic dependence.

Nevertheless, I do briefly consider the idea that some of the dependence relations in B&R's minimal models can be given a causal interpretation. I do this simply because I do not think a causal interpretation is as obviously wrong as B&R imply. A causal interpretation is more plausible for some common features than others, though I do not commit myself here to a causal interpretation of any of them.

On my account, RG plays a central role in discovering explanatorily relevant features and demonstrating that they are relevant (Section 4.3 shows how). This makes RG not a kind of explanation distinct from common features explanation, but an essential method scientists use to construct common features explanations.

4.2. B&R's Minimal Models

B&R present two minimal models whose explanatoriness they argue cannot be captured by a common features approach. These are the Lattice Gas Automaton (LGA) model of fluid dynamics and Fisher's optimality model of 1:1 sex ratios.

LGA accurately predicts macroscopic fluid behavior that is described by the Navier-Stokes equations ("Navier-Stokes behavior," for short). The model consists of a hexagonal lattice on which each particle has a lattice position and one of six directions of motion (momentum vectors). Each particle moves one step in its direction of motion and if some "collide", so that

their total momentum adds to zero, then those particles' directions of motion rotate 60°. With thousands of particles and steps, and some smoothing out of the data, an overall pattern of motion emerges that is incredibly similar to real fluid motion (Goldenfeld and Kadanoff 1999, 87).

The second model presented by B&R is Fisher's model of the 1:1 sex ratio. The biological question that Fisher's (1930, 141–3) model was designed to answer is why population sex ratios are often 1:1. Hamilton (1967) provides a succinct summary of Fisher's argument. If males are less common than females in a population, then a newborn male has better mating prospects than a newborn female. In this situation, parents genetically disposed to have male offspring will tend to have more than the average number of grandchildren. This will cause the genes for the tendency to have male offspring to spread. As male births become more common and a 1:1 sex ratio nears, the advantage of the tendency to produce males disappears. Since the same reasoning holds if females are the more common sex, 1:1 is the equilibrium sex ratio (Hamilton 1967, 477).

If, then, male and female offspring cost the same amount of resources on average, a 1:1 sex ratio will result. More generally, any sex ratio can be calculated as $C_M / (C_M + C_F)$, where C_M is the average resource cost of one male offspring and C_F is the average resource cost of one female offspring (B&R, 367).

4.3. B&R's Account of the Explanatoriness of Minimal Models

B&R's account of the explanatoriness of their minimal models makes use of the concepts of the renormalization group (RG) and universality classes. Here I explain these concepts and how they fit into B&R's account.

RG is a method of coarse-graining, reducing degrees of freedom or the number of details. B&R (362) discuss one such procedure: Kadanoff's block spin transformation. Consider a lattice of particles, each with an up or down spin. Group the spins into blocks of, for example, four spins and average over each block. One averaging procedure is called 'majority rule', in which a block of four spins is replaced by the most common spin in the block. If there is no most common spin, choose one randomly (see McComb 2004). This reduces the number of spins in the lattice by a factor of four. The length between spins, or the lattice constant, is greater after averaging, so it is then rescaled to the old lattice constant. Near a critical point, the length across which spins are correlated, or the correlation length, increases and eventually diverges to infinity. When this is the case, averaging over correlated blocks of spins and then rescaling the lattice preserves the macroscopic behavior of the lattice with fewer degrees of freedom (microscropic details) (Huang 1987, 441–2). The irrelevant details are thereby eliminated.

With the concept of RG in hand, we can define a universality class. After repeated application of RG, certain systems will reach the same fixed point, a state at which RG no longer has an effect. The class of all systems that will reach the same fixed point after repeated application of RG is a universality class.

Using RG, it can be discovered that all systems exhibiting Navier-Stokes behavior, including LGA, form a universality class that shares the following three features:

> 1. Locality: A fluid contains many particles in motion, each of which is influenced only by other particles in its immediate neighborhood.
> 2. Conservation: The number of particles and the total momentum of the fluid is conserved over time.
> 3. Symmetry: A fluid is isotropic and rotationally invariant. (B&R 360; from Goldenfeld and Kadanoff [1999], 87)

Similarly, an RG-type story would show that all populations exhibiting a 1:1 sex ratio, including

Fisher's model, form a universality class and share the feature of linear substitution cost, that is, the average resource cost of male offspring is equal to the average resource cost of female offspring.

According to B&R, although RG demonstrates that diverse systems share features with their minimal models, it is not this fact that accounts for the explanatoriness of their minimal models. An account of why minimal models are explanatory must, according to them, answer the Three Questions presented above. B&R argue that RG answers Q2, for both LGA and Fisher's model, because the RG transformation eliminates details that are irrelevant. They write, "By performing this [RG] operation repeatedly, one can answer question Q2 because the transformation in effect eliminates details or degrees of freedom that are irrelevant" (362). However, RG alone does not answer this. Q2 asks why the heterogeneous details are irrelevant and RG only shows us that the details are irrelevant. The answer appears to be, "The details are irrelevant because, as RG shows, the same macro-behavior results no matter the details." But this is uninformative.[40]

RG is also supposed to answer Q3 by demonstrating that all the fluids within LGA's universality class share the common features of locality, conservation, and symmetry, and that all populations in Fisher's model's universality class share linear substitution cost (363, 372). B&R write that,

> A derivative, or *by-product*, of this [RG] analysis is the identification of the shared features of the class of systems. In this case, the by-product is a realization that all the systems within the universality class share the common features locality, conservation,

[40] An anonymous referee suggests the possibility that in this case there is no clear distinction between showing why and showing that the details are irrelevant. I agree that in the LGA case the distinction seems blurry. However, there are clear cases. For example, the entire cerebellum appears to be irrelevant to consciousness, even though it contains more neurons than the cerebral cortex. Knowing this does not tell one why the cerebellum is irrelevant – according to one popular theory, it has to do with the cerebellum's lack of informational integration (Tononi and Koch 2015).

> and symmetry. Thus, we get an explanation of why these are the common features as a by-product of the mathematical delimitation of the universality class. (363; their emphasis)

The by-product is merely the identification of the shared features, not why they are shared. Again, RG merely shows that these features are shared across diverse systems, not why they are shared. Perhaps B&R's suggestion is that the fact that RG demonstrates that the details are irrelevant explains why the common features are shared. But this boils down to, "These features are shared across diverse systems because no other features are shared." This is also uninformative. RG alone does not explain why locality, symmetry, and conservation are present in, for example, water and LGA, but not anisotropic liquid crystals. Answering that question requires investigation of specific fluids. One reason why liquid crystals are not in the same universality class as LGA and water is that their often rod-shaped particles result in directional preference and lack of symmetry (Priestley *et al.* 1975). Liquid crystals therefore cannot be accurately modeled using the unmodified Navier-Stokes equations. The addition of a stress tensor or coupling with a Q-tensor system is required to take into account the anisotropy of liquid crystals (Badia *et al.* 2011; Paicu and Zarnescu 2012). Similarly for Fisher's model: RG alone does not explain why the average resource cost of male and female offspring is equal in, for example, sheep, mule deer, and so on, but not in, for example, bees.

Finally, the answer to Q1 follows from the answers to Q2 and Q3. Obviously, if B&R are mistaken about their answers to Q2 and Q3, then they are also mistaken about Q1.

Perhaps I have interpreted B&R too narrowly, and they do not mean that RG alone can answer their Three Questions. If I am right about RG, B&R are wrong merely about how to go about answering the Three Questions, not that answers are required. Next, then, I present two

arguments that such a story is not required, that answering their Three Questions is unnecessary for an account of the explanatoriness of LGA and Fisher's model.

The first argument rests on an analogy with a commonplace case of multiple realizability. Batterman (2000; 2002b, §5.5) has plausibly argued that universality just is multiple realizability:

> That microstructurally different systems fall in the same universality or equivalence class, is the physicists' way of saying that the upper level universal behavior is multiply realized. And so, the explanation of the universality of the behavior is an explanation of the behavior's multiple realizability. (2000, 129)

The diverse systems in a universality class multiply realize some universal behavior. Therefore, Batterman argues, RG or similar methods can explain cases of multiple realizability. The following analogy, then, is apt, and the lessons derived therefrom should generalize to B&R's account of LGA and Fisher's model. If the lessons do not generalize, B&R need to explain why.

Diverse fluids exhibit similar behavior (for example, critical behavior) under certain conditions (for example, near critical points). Similarly, diverse objects, such as apples, tomatoes, and bowling balls, exhibit similar behavior (for example, rolling) under certain conditions (for example, on an incline plane[41]). Rolling under these conditions is universal, or multiply realizable, in apples, tomatoes, and bowling balls; apples, tomatoes, and bowling balls are in the same universality class with respect to rolling. We would like to know why this is; why apples, tomatoes, and bowling balls all roll on an incline plane. These diverse objects behave similarly in certain conditions in virtue of possessing a similar property, (approximate) sphericity. It is their (approximate) sphericity that disposes them all to roll when placed on an incline plane. That fact could be discovered by some RG-like method. That they all share the

[41] And in a suitable gravitational environment and so on.

relevant property of sphericity and that all of their other properties, such as size[42] and color, are irrelevant to rolling on an incline plane is what explains this similar behavior and allows us to answer w-questions about it. A minimal model of spherical objects would be in the same universality class as apples, tomatoes, and bowling balls, and would explain their similar behavior in certain conditions in virtue of accurately representing the relevant property, (approximate) sphericity. Why should our account of the explanatoriness of B&R's minimal models differ from this one?

The further question – Why are the remaining heterogeneous details, such as the size, material, and color of these objects, irrelevant for the disposition to roll? – which is analogous to B&R's Q2, is unnecessary for an account of the explanatoriness of our minimal model of spherical objects. Why, for example, the color of an object does not matter to its rolling on an incline plane is a question that can only be answered by a physical investigation into the dispositions bestowed by color. An investigation in color physics would reveal why the disposition to roll on an incline plane is not one of the dispositions bestowed by color. Such an investigation would be unnecessary for knowing or showing that color is irrelevant to the disposition to roll and, therefore, unnecessary for an account of the explanatoriness of our minimal model of rolling.

The question analogous to B&R's Q3 is, "Why do very different objects, such as apples, tomatoes, and bowling balls, all have sphericity in common?" Intuitively, an answer to this question is beside the point to answering the question of why these objects behave similarly in certain conditions, why they all roll when placed on an incline plane. Furthermore, this question

[42] Obviously there are limits in the example as described. For example, if the size of the bowling ball (or apple or tomato) were too large, it would crush the incline plane, unless the plane is sufficiently strong. Assume all these deviant cases are excluded.

seems to have no good answer. Yet the absence of an answer does not suggest that there is no explanation of these diverse objects' disposition to roll on an incline plane. Similarly, there may be no good answer to the question of why some diverse fluids share locality, conservation, and symmetry, or why some diverse populations share linear substitution cost. The story about why large classes of features are irrelevant that is required by B&R may not be available. This analogy should motivate the claim that such a story is unnecessary to answer the question of what makes LGA and Fisher's model explanatory. B&R need to say why answers to the Three Questions are necessary in the cases of LGA and Fisher's model, but not in my rolling case or similar cases of multiple realizability.

The above analogy is entirely consistent with Batterman's own remarks on the multiple realizability of pain:

> Suppose that physics tells us that the physical parameters α and γ are the (only) relevant physical parameters for the pain universality class. That is, that N_h, N_r, and N_m have these features in common when certain generalizations or regularities about pain are the manifest behaviors of interest observed in each of humans, reptiles, and martians. Equivalently, physics has told us that all the other micro-details that legitimately let us think of N_h, N_r, and N_m as heterogeneous are irrelevant. We then have our explanation of pain's realizability by wildly diverse realizers. (2000, 133; see also 2002b, §5.5)

This appears to be a common features explanation of exactly the type given above for the multiple realizability of rolling on an incline plane. N_h, N_r, and N_m are the realizers of pain in humans, reptiles, and martians, respectively. They are all in the pain universality class. An RG-type procedure might discover that α and γ are the only relevant common features shared by these realizers. This would be enough to explain the multiple realizability of pain in humans, reptiles, and martians. Further questions such as why humans, reptiles, martians, sentient robots, and everything else in the pain universality class have the pain-conferring features α and γ in

common may have no good answer. Answers to the Three Questions are therefore unnecessary for an explanation of the multiple realizability of pain.

There is another reason why answering the Three Questions is unnecessary. Were answers necessary for an account of the explanatoriness of LGA and Fisher's model, a regress would loom. They write, “Simply to cite locality, conservation, and symmetry as being explanatorily relevant actually raises the question of why those features are the common features among fluids” (361). Similarly,

> Common features accounts would likely cite the fact that the different fluids have locality, conservation, and symmetry in common as explanatorily relevant and maybe even as explanatorily sufficient. However, as we emphasized in section 3.3, this is a mistake. The fact that the different fluids all possess these common features is also something that requires explanation. (374)

Common features are insufficient to explain macroscopic fluid behavior because, B&R argue, they do not answer the further question of why these features are common. With respect to 1:1 sex ratios, B&R write,

> Were we simply to cite the fact that all these populations have the common feature of linear substitution cost, we would fail to explain this universal behavior. The reason for this is that we can equally well ask why the populations of different species distinguished by different mating strategies, and so on, all exhibit a linear substitution cost and why they display the 1:1 sex ratio. (374)

This appears to be an injunction against explanations that appeal to things that also require explanation.[43] But if it is a mistake to explain something by appeal to something else that requires explanation, then nearly all explanations are mistaken. B&R need to explain why the chain of explanation should stop where they say it should.

To conclude this section, I have found two problems with B&R's account of the explanatoriness of their minimal models. First, it does not appear that RG alone can answer the

[43] This point is also made by Lange (2015, 303–4).

Three Questions. Perhaps they did not mean to imply as much. The second problem is that answering the Three Questions is unnecessary. I gave two arguments for this. First, it is plausible that answers to analogous questions in similar cases of multiple realizability are unnecessary (and potentially unavailable, without thereby threatening explanation), and, second, were answers to the Three Questions necessary, a regress would loom. Having argued against B&R's account, I now present my own common features account.[44]

4.4. Generalizing the Ontic Conception

The account I propose is similar to the accounts proposed by Bokulich (2011), Rice himself (2013), and Saatsi and Pexton (2013), though I give my account an ontic spin[45]. These authors follow Woodward (2003) in requiring that an explanation answer what-if-things-had-been-different questions ("w-questions"). According to Woodward, an explanation "must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways" (2003, 11). This requires the accurate representation of the objective relations of dependence between the explanandum phenomenon and the features on which it depends.

Woodward is explicit that it is in virtue of conveying counterfactual information that causal claims are explanatory (2003, 210–220). Since noncausal dependence relations can also convey counterfactual information, they can, therefore, also be explanatory[46]. For example,

---

[44] Perhaps it will be said that I have missed the distinctive feature of Fisher's model: that it is an equilibrium explanation. According to Sober (1983), "Where causal explanation shows how the event to be explained was in fact produced, equilibrium explanation shows how the event would have occurred regardless of which of a variety of causal scenarios actually transpired" (202). Equilibrium explanations show how many of the causal details are irrelevant to the explanandum. This presents no challenges I have not already discussed here at length. The common features account given here is much like Rice's (2013) own account of equilibrium explanation.

[45] See also Ruben's (1990) 'realist' account of explanation that emphasizes determinative and dependency relations and Thalos' (2002) discussion of causal dependence as only one form of explanatory dependence.

[46] Woodward (2003, §5.9) himself suggests dropping the causal requirement in certain cases where an interventionist interpretation is implausible. See also Strevens (2008, 177–80).

Saatsi and Pexton (2013) present an explanation of Kleiber's law, an allometric scaling law that relates an organism's body mass to a biological observable (West *et al.* 1999). The precise details of the explanation are irrelevant for our purposes. What matters here is that there is a feature, the scaling exponent, that counterfactually depends on the dimensionality of the organism. It is plausible that this counterfactual dependence relation contributes explanatory power, yet it is implausible that the dimensionality of organisms is a causal variable that can, in practice or in theory, be intervened upon (Saatsi and Pexton 2013, 620).

Salmon (1984, 1989) distinguished between epistemic, modal, and ontic conceptions of explanation. These are conceptions of what a scientific explanation aims to show of the explanandum phenomenon: that it is expected to occur, that it had to occur, and that it fits "into a discernible pattern," respectively (1984, 121). For Salmon, the "discernible pattern" into which the explanandum phenomenon is fit is structured by causal processes, causal interactions, and causal laws (1984, 132). "[W]e explain," wrote Salmon, "by providing information about these patterns that reveals how the explanandum-events fit in" (1989, 121). Explanation is not about nomic expectability or nomic necessity, but about fitting the explanandum into "discernible patterns," "relationships that exist in the world" (1984, 121). This need not be construed solely causally – it is a mistake to equate the ontic conception with the causal-mechanical account of explanation. Salmon actually did not think causation was essential to the ontic conception:

> It could fairly be said, I believe, that mechanistic explanations tell us how the world works. These explanations are local in the sense that they show us how particular occurrences come about; they explain particular phenomena in terms of collections of particular causal processes and interactions – or, perhaps, in terms of noncausal mechanisms, if there are such things. (1989, 184)[47]

[47] See also, for example, "[T]he ontic conception focuses upon the fitting of events into natural regularities. Those regularities are *sometimes*, if not always, causal" (1989, 120; my emphasis) and, "explanations reveal the mechanisms, causal *or other*, that produce the facts we are trying to explain" (1989, 121; my emphasis). Salmon then says that Railton's (1978; 1981) account is an ontic conception even though, "His view is more lenient than

For Salmon, what was essential to the ontic conception was that, "the explanation of events consists of fitting them into the patterns that exist in the objective world" (1989, 121). We can and should hold on to the ontic conception while accepting many of the criticisms and limitations of causal explanation, including those provided by B&R. There are noncausal dependence relations in which an explanandum phenomenon can stand to other worldly items. Explanation remains, then, a matter of fitting the explanandum phenomenon into "discernible patterns" and "relationships that exist in the world," all while acknowledging that these worldly patterns and relationships can be noncausal.

The ontic-epistemic debate has shifted twice since Salmon (Illari 2013). Salmon framed the debate in terms of what explanations do. After Salmon, the debate was framed metaphysically, as a debate about what explanations are. The ontic conception was associated with the claim that scientific explanations are (almost always causal) dependence relations in the world; the epistemic conception was associated with the claim that scientific explanations are epistemic states or representations. Craver's (2014) most recent formulation of the ontic conception backs away from the metaphysical claim that explanations are ontic structures in the world and focuses on demarcatory and normative constraints on explanation.[48] Craver (2014) writes that according to the ontic conception, "in order to satisfy these two objectives [of explanatory demarcation and explanatory normativity], one must look beyond representational structures to the ontic structures in the world" (28). That is, attention to ontic structures, rather

mine with regard to noncausal explanation" (1989, 121). Salmon also clearly thought that laws, construed as *ontic regularities*, can be explanatory (see, for example, 1984, 17–8, 121; 1989, 120). See especially (1989, 120, 129) for explicit claims that focus on the laws themselves, rather than law-*statements*, leads to the ontic conception.

[48] According to Illari (2013, 241), Craver (in personal communication with Illari) holds that this has always been the debate.

than epistemic or representational form, is required in order to demarcate explanation from other scientific achievements, like prediction, and to distinguish good from bad explanations, how-possibly from how-actually explanations, and explanatorily relevant from irrelevant features (2014, 51).[49]

The generalized ontic conception, then, is an ontic conception because it embraces Craver's claim that achieving the objectives of explanatory demarcation and normativity requires attention to the ontic. It is generalized because it says that attention to more of the ontic than just the causal-mechanical is required to achieve those objectives – attention to all ontic structures on which the explanandum depends is required[50].

The ontic conception, unhindered by a strictly causal-mechanical interpretation, retains the ability to demarcate explanation from description and prediction. Explanations provide information about relations of ontic dependence, causal and noncausal, which can be used to answer w-questions about the explanandum phenomenon. Understanding is possessing this information, and, therefore, knowing answers to w-questions[51]. Norms of explanation immediately fall out of this account: The more relevant dependencies that are represented for a given phenomenon and the more irrelevant dependencies that are not, the more w-questions can be answered, the better the explanation of that phenomenon.

Let me clarify the relation between the aspect of my account that emphasizes dependence relations and the counterfactual aspect that emphasizes the ability to answer w-questions. These aspects are tightly intertwined, but relations of dependence are not "analyzed" in terms of

[49] Under this framing of the debate, Wright (2012) overemphasizes the role that lexical ambiguity plays in the case for the ontic conception. The argument, which I do not have space here to defend, for Craver's claims about explanatory demarcation and normativity does not require any lexical ambiguity of the term 'explanation'.

[50] I thank an anonymous referee for pressing me to clarify the ontic conception and my account's relation to it.

[51] More needs to be said about understanding than I am able to say here. See, for example, Strevens (2013) for the kind of view to which I am sympathetic.

counterfactuals or 'reduced' to counterfactuals. Analysis and reduction apply to terms, concepts, or theories, not the things to which they refer. Rather, relations of counterfactual dependence hold in virtue of, or are grounded in, relations of ontic dependence. Like supervenience, counterfactual dependence is a modal concept (Heil 2003, 37). Different relations of ontic dependence could ground supervenience, including, among others, identity, constitution, and causal sufficiency (2003, 67). Supposing that what grounds counterfactual dependence relations also makes (descriptions of) them true, we can put this in terms of truthmakers: relations of ontic dependence provide truthmakers for counterfactuals.[52]

It is only with information about dependence that one can answer w-questions. This is why the ontic aspect of my account is inseparable from the counterfactual aspect. This is why one cannot say that explanation is a matter of answering w-questions, but not a matter of accurately representing dependencies. Bokulich (2011), Rice (2013), Saatsi and Pexton (2013) emphasize the importance for explanation of the ability to answer w-questions and are silent about ontic relations, but these issues cannot be separated. Consider the counterfactual, "If population P had lacked linear substitution cost, it would not have a 1:1 sex ratio." What grounds this counterfactual is the (perhaps causal) dependence between the population's linear substitution cost and its 1:1 sex ratio. Those who think of explanation in terms of the ability to answer w-questions should therefore embrace the account presented here.

The ontic aspect of my account also allows one to distinguish explanatorily relevant from irrelevant counterfactuals. The length of a flagpole's shadow can be derived from the height of the pole and the angle of elevation of the sun (Bromberger 1966). This derivation is symmetric.

[52] Though I think this way of putting it is illuminating, it is controversial both in light of possible-world semantics for counterfactuals and in light of disagreement about the relation between grounding and truthmaking. For a survey of possible relations between grounding and truthmaking, see Griffith (2014).

That is, one can also derive the height of the flagpole from the length of a flagpole's shadow and the angle of elevation of the sun. It seems, then, that if the shadow had been longer and the sun in the same position, then the flagpole would have been higher. Yet it does not seem true that this explains the height of the flagpole. Here it is plausible that the explanatory asymmetry is provided by causal asymmetry: the derivation of the length of the pole's shadow counts as explanatory because that derivation, but not the reverse derivation, tracks causes (Hausman 1998; Woodward 2003). This lesson can be generalized to cases of noncausal dependence: in general, when there are explanatory asymmetries, these are due to asymmetries in ontic dependence.

Symmetry provides a nice example of something on which fluid behavior noncausally depends. As I mentioned above, there are fluids, like anisotropic liquid crystals, that have a preferential alignment due to their banana- and rod-shaped molecules and therefore cannot be accurately modeled using the unmodified Navier-Stokes equations. The dependence of the macro-behavior of liquid crystals on the shape of their particles is plausibly not a causal dependence or mechanistic dependence. A feature or disposition of the whole liquid, its macro-behavior, depends on the features of its mereological parts, so construing this dependence causally is inappropriate (assuming, plausibly, that parts and wholes cannot stand in causal relations to each other; see Craver and Bechtel 2007). Yet, it is also plausible that the particles are not a mechanism that produces, underlies, or maintains the fluid's macro-behavior. Mechanisms are organized in a way mere aggregates are not (Craver 2001), and, while I recognize that there is something of a continuum here, fluid particles do not appear to have the requisite organization to constitute a mechanism. Here, then, is an instance of ontic dependence

that is neither causal nor mechanistic, but is asymmetric and can be used to answer w-questions about fluid behavior.[53]

B&R remark only in passing that it "stretches the imagination" to think of locality, symmetry, conservation as causally relevant (360)[54]. I agree, but I do think it is plausible that linear substitution cost can be given a causal interpretation, though I do not think a causal interpretation is required for that feature to be explanatory. Woodard (2003) has given the most influential account of causal relevance. Very briefly, according to Woodward, *x* is causally relevant to *y* if and only if a sufficiently surgical manipulation (or "intervention") of *x* would change *y*. Here, 'sufficiently surgical' means that a manipulation of *x* that would change *y* would do so only via the pathway from *x* to *y*.

It is important to note that on Woodward's view, the manipulation need not be physically possible. All that is necessary is that relevant scientific theory be able to answer what would happen under the imagined intervention. For example, considering the counterfactual claim that changes in the position of the moon cause changes in the motion of the tides, Woodward writes,

> Newtonian theory and familiar rules about the composition of forces tell us how to subtract out any direct influence from such a process so that we can calculate just what the effect of, say, doubling of the moon's orbit (and no other changes) would be on the tides, even though it also may be true that there is no way of actually realizing this effect alone. In other words, Newtonian theory itself delivers a determinate answer to questions about what would happen to the tides under an intervention that doubles the moon's orbit, and this is enough for counterfactual claims about what would happen under such interventions to be legitimate and to allow us to assess their truth. (2003, 131)

If physical theories and biological theories can tell us what would happen under hypothetical interventions, then causal relevance can be established.

---

[53] I suspect that many explanations of dispositions in terms of their micro-bases will have this noncausal, non-mechanistic structure.

[54] Lange (2015, 300) points out that this is plausible if it means that locality, symmetry, and conservation are not causes, but implausible if it means that they cannot figure in causal explanations. See also fn. 55 below.

A causal interpretation of linear substitution cost is plausible on a manipulationist account. Recall that linear substitution cost is equality between the average resource costs of male and female offspring. Here is a hypothetical intervention on average resource cost: inject all and only the males of a population with a fluid that has the only effect of raising their metabolism and increasing their average resource cost. Do this over many generations in a population that initially had a 1:1 sex ratio and you will eventually see a deviation from a 1:1 sex ratio.

One might object that this hypothetical intervention does not show that linear substitution cost is causally relevant to 1:1 sex ratios, only that metabolism is causally relevant, since this is what was manipulated. This objection is conceptually confused. In the case at hand, manipulating metabolism just is manipulating average resource cost. It does not matter if manipulating metabolism were but one way among many of manipulating average resource cost. There are usually many different ways to manipulate a variable. Although, according to the generalized ontic conception, linear substitution cost need not be causally relevant to be explanatorily relevant, it plausibly is causally relevant on the manipulationist account.

It is much less plausible that conservation and locality are causally relevant to the macro-behavior of fluids. Conservation is a paradigm law of nature. It is hard to imagine any hypothetical interventions that would alter this regularity. One can imagine “local miracles,” local speedings up, slowings down, and poppings into and out of existence of a fluid's particles, and this would certainly change the macro-behavior of the fluid. Physical theory might even be able tell us what would happen in such a contranomic or counterlegal scenario, but it is highly implausible to construe laws as causally relevant in the interventionist sense because laws are not

events or objects and particles are mereological parts of the fluid.[55]

According to the generalized ontic conception, then, LGA explains Navier-Stokes behavior and Fisher's model explains 1:1 sex ratios in virtue of accurately representing all and only the relevant features: symmetry, locality, and conservation for fluid behavior, and linear substitution cost for 1:1 sex ratios. Knowing that these features alone are the relevant ones allows one to answer w-questions about fluid behavior and 1:1 sex ratios. The essential role RG plays is in discovering and demonstrating that these are the relevant features. RG and universality classes do not provide a kind of explanation distinct from common features explanations. Rather, RG and similar procedures are necessary methods used in the construction of common features explanations.[56][57]

[55] This is not to deny that conservation laws are causally relevant in the sense that they govern or constrain all *causal interactions*, in Salmon's (1984, 169–70) sense of that term. Nor am I denying that citing a law can provide information about a phenomenon's causal history (Skow 2014). I am only denying that conservation laws are causally relevant in the interventionist sense. See Lange (2007; 2011) for valuable discussions of the nature and explanatory status of conservation laws.

[56] Cf. Batterman (2000): "The RG type analysis illuminates those physical features that *are* relevant for the upper level universal behavior, *and at the same time demonstrates that all of the other details which distinguish the systems from one another are irrelevant*" (128; original emphasis). More compactly, "[RG] is a *method for extracting structures* that are, with respect to the behavior of interest, detail independent" (2000, 128; added emphasis). Also, B&R: "[T]here are a number of techniques for *demonstrating that* a large class of details of particular systems is irrelevant to their macroscale behavior" (371; added emphasis). These quotations are consistent with my account of the role of RG.

[57] Reutlinger (2014) has argued that RG explanations are noncausal because they are a kind of "distinctively mathematical explanation," although they do not exploit mathematical necessity, in contrast to Lange's (2013) account of distinctively mathematical explanation. Rather, the mathematical *operations* involved in RG account for RG's explanatory power. Reutlinger writes,

> The mathematical explanatory power is derived from... the [RG] transformations and flow of Hamiltonians [to a fixed point]. […] Both the transformations and the 'flow' are mathematical operations, which, ultimately, *serve the purpose to reveal something that two fluids have in common* despite the fact that their "real physical" Hamiltonians (or "initial physical manifolds") are strikingly different. (1166, 1168; my emphasis)

I agree that the mathematical operations of RG reveal common features, but I do not agree that those operations are the sole contributors of explanatory power. If that were true, we would seem to have a case where representing the things on which an explanandum depends does not contribute explanatory power, but the *method(s)* used to reveal, discover, or demonstrate the relevance of those things does. This seems false in my multiple realizability of rolling example, in which case it cannot be true of explanations of multiple realizability in general. That is, it seems false that representing their shared (approximate) sphericity does *not* contribute explanatory power to the explanation of the multiple realizability of rolling by apples, tomatoes, and bowling

4.5. Conclusion

Batterman and Rice are at the forefront of a philosophical exploration of the limits of causal explanation. They have argued forcefully and plausibly that certain models in physics and biology are not explanatory in virtue of accurately representing causes (for example, Batterman 2002a; 2002b; Rice 2012; 2013). In their recent paper (Batterman an Rice 2014), they use the minimal models to critique the explanatory requirement of accurate representation, regardless of whether the features accurately represented are causal.

According to B&R, the explanatoriness of LGA and Fisher's model is captured by a story about why heterogeneous details are irrelevant, a story that answers the Three Questions. I identified two problems with this account. First, RG alone cannot answer the Three Questions. Perhaps RG in conjunction with other methods can. Even so, the second problem is that answers to these questions are in fact unnecessary. I argued for this by showing 1) that answers to analogous questions in an analogous case of multiple realizability are unnecessary, and 2) that if answers to the Three Questions were necessary, a regress would loom.

B&R have rightly stressed the significance of RG explanation, but have misplaced where that significance lies. These methods do not provide novel kinds of explanation. RG is a unique method that is necessary to extract the relevant features of the world that explain the phenomena in which physicists are often interested. The explanatoriness of the minimal models they present,

---

balls, but that the method(s) used to reveal, discover, or demonstrate that (approximate) sphericity is the only relevant, common property *does* contribute. Similarly for the multiple realizability of pain, briefly discussed above. Rather, representing the only relevant common features on which our explanandum depends is what contributes explanatory power, by allowing us to answer w-questions about it. The methods, mathematical or not, that we use to *discover* that (approximate) sphericity is the only relevant property do not contribute any explanatory power in themselves; they are simply tools used in the construction of the "common features" explanation. This is how I see the role of RG. Note that if Reutlinger's distinctively mathematical account is to be extended to other minimal models, some analogues of the mathematical operations of RG must be specified, since those are the operations that he argues contribute explanatory power. In, for example, biological contexts, it is unclear what such operations could be.

LGA and Fisher's model, can be adequately captured by a common features approach, the generalized ontic conception. These minimal models explain by accurately representing the features on which their explananda depend, causally or noncausally. These accurate representations can then be used to answer w-questions about the explananda, which contributes to their explanatory power.

Chapter 5. Information and Explanation: A Dilemma

I identify a dilemma for anyone who holds a theory according to which explanation is the conveyance of some kind of information (e.g., Lewis 1986; Railton 1981). I also show how an informational theory of explanation is implicit in core arguments of mechanists (Piccinini and Craver 2011; Zednik 2015). However, informational theories seem to conflict with some lay and scientific commonsense judgments and a goal of the theory of explanation, because information is relative to the background knowledge of agents (Dretske 1981). Sometimes a model is an explanation *simpliciter*, not just an explanation relative to some particular agent. We would also like a philosophical theory to tell us *when* a model is an explanation *simpliciter*, not just when a model is an explanation relative to some particular agent. I sketch a solution by distinguishing explanation *simpliciter* from explanation-to and relativizing the former to what I call "total scientific background knowledge".

5.1. Introduction

Venerable philosophical theories of explanation have identified explanation with conveying information of some kind (e.g., Jackson and Pettit 1990; Lewis 1986; Railton 1981; Skow 2014). I will call these informational theories of explanation. Usually informational theories are causal, according to which explanations provide information about the explanandum phenomenon's causal history, but they need not be. There have been innumerable critiques of such causal theories (e.g., Batterman and Rice 2014; Sober 1983; for responses see Povich forthcoming a and Skow 2014), but none, as far as I know, targets the apparent counterintuitive consequences of their reliance on information.

Informational theories seem to conflict with some commonsense lay and scientific

judgments and goals of the theory of explanation, because information is relative[58] to the background knowledge of agents (Dretske 1981). For example, it seems that sometimes a model is an explanation *simpliciter*, not just an explanation relative to some particular agent. We would also like a philosophical theory to tell us *when* a model is an explanation *simpliciter*, not just when a model is an explanation relative to some particular agent. However, relying on information seems to give us only the latter. I sketch a solution to this dilemma[59] by distinguishing scientific explanation *simpliciter* from explanation-to and relativizing information in scientific explanations *simpliciter* to what I call "total scientific background knowledge" (TSBK).

In Section 5.2, I sketch a typical informational theory of explanation, so that we have a clear example of such a theory, and provide textual evidence that such a theory has actually been held by prominent philosophers.

In Section 5.3, I show how an informational theory of mechanistic explanation is implicit in core arguments of the mechanistic research program. Specifically, an informational theory is implied by Piccinini and Craver's (2011) argument that functional analyses are mechanism sketches and at least implies the crucial premise in Zednik's (2011, 2015) argument that

[58] An obvious way out of the dilemma is to deny the relativity of information. At least one account of information denies this relativity, but that account is controversial (Cohen and Meskin 2006; for objections, see Demir 2008 and Scarantino 2008; for a reply, see Meskin and Cohen 2008). For the purposes of this paper I will assume that information is relative.

[59] I have decided to structure this article as a dilemma for informational theorists of explanation because I do not present much of an *argument* for an informational theory here (see Lewis 1986; Povich forthcoming a; Skow 2014). Of course, the article could be seen as gesturing towards a general inconsistent triad between an informational theory of explanation, the agent-relativity of information, and a kind of objectivity of explanation (as demonstrated by lay and scientific commonsense judgments of explanatoriness). One point in favor of an informational theory is that information about causal, mechanistic or other kinds of dependence is all that is needed to answer what-if-things-had-been-different questions about the explanandum phenomenon, which many otherwise different philosophers think is essential to explanation (e.g., Chirimuuta 2014; Craver 2007; Povich forthcoming a; Rice 2015; Woodward 2003).

dynamical systems models can be mechanistic explanations[60]. This makes a solution to the dilemma even more pressing.

In Section 5.4, I explain why information is relative to background knowledge and illustrate this with Dretske's (1981, 78) example of the shell game. I can then spell out more clearly why informational theories of explanation have the counterintuitive consequences they do.

In Section 5.5, I argue that informational theories can avoid the counterintuitive consequences by distinguishing scientific explanation *simpliciter* from explanation-to (so-called because it is an explanation *to* a particular agent). Information in scientific explanation *simpliciter* is relativized to what I call "total scientific background knowledge" (TSBK), the current total store of propositions that are known by scientists[61]. Information in explanation-to is relativized to the background knowledge of the particular agent to whom the explanation is given.

5.2. An informational theory of explanation

In this section, I sketch a typical informational theory of explanation (inspired by Lewis 1986), so that we have something concrete with which to work.

> **The Informational Theory of Explanation (ITE)**: A model is an explanation when and only when it provides information about the explanandum phenomenon's causal history.

This is a causal version of an informational theory since the kind of information that is deemed explanatory is information about causes. Other versions are certainly available, to which my dilemma and solution will apply. For example, a mechanistic version would say that a model is

[60] Hochstein (2016) makes a somewhat similar argument, though he is not concerned with information. He argues that implicit in mechanists' arguments is the idea that any representation *of* a mechanism is a mechanistic explanation, regardless of the form of the representation.

[61] Anti-realists can replace "knowledge" with their preferred substitute. Ditto for "veridical information" in Section 2.

an explanation when and only when it provides information about the explanandum phenomenon's mechanism. Here I use the causal version as a typical example; I return to the mechanistic version in the next section. All the points that follow apply, *mutatis mutandis*, to all versions of **ITE**. For my purposes, I need not commit to any particular theory of causation.

To provide information, in the intended sense, about an explanandum phenomenon's causal history is to reduce uncertainty about its causal history or to reduce the space of possible causal histories responsible for it (Dretske 1981; Skow 2014). A causal history is merely the explanandum phenomenon's entire causal chain or network (Lewis 1986). To provide information about an explanandum phenomenon's causal history, it is not necessary to reduce the space of possible causal histories to one, which would be to provide "complete" information, in an intuitive sense. Also, to provide complete information in this sense is not necessarily the same as providing a complete *explanation*. Due to the (often pragmatic) norms governing completeness of explanation, a complete explanation could be provided without complete information. I have made **ITE** non-comparative so that I will not have to address norms of completeness (Kaplan and Craver unpublished). I take explanatory information to be veridical.[62]

I include "model" in **ITE** because when we make judgments about explanations, these explanations are usually embodied in models. Similarly, we would like a philosophical theory to tell us when something is an explanation, and that something is usually a model.[63] By "model" or "representation," which I use synonymously, I mean a structure (e.g., a concrete replica, a mathematical equation, a diagram, or a linguistic description) that is interpreted to represent a

[62] I do not take much to hang on this (cf. Lewis 1986, 226) and I will not address idealization here (see Craver 2014; Weisberg 2013). This should not affect my argument or solution, for information is still relative to background knowledge, whether or not it is veridical.

[63] Even proponents of the ontic conception recognize both that the term "explanation" is used this way and that their theory needs to address these questions (Craver 2014; Craver and Kaplan unpublished).

target system (Weisberg 2013). Although I am ignoring Weisberg's (2013) distinction between models and model descriptions, this should not affect the points that follow.

It might be thought that information is too weak a basis for explanation (Carl Craver personal communication). Part of the motivation for causal-mechanical accounts was to avoid the verdicts on which the covering law model floundered. One kind of case was where merely informational relations were counted as explanatory. For example, the barometer provides information about the storm, but does not explain the storm (Salmon 1989)[64]. However, **ITE** does not entail that the barometer explains[65] the storm, because, although a barometer reading reduces uncertainty about the occurrence of a storm, it does not reduce uncertainty about the causes of the storm. Nor does the fact that barometer readings and storms are correlated reduce uncertainty about the causes of the storm (barring application of Reichenbach's controversial Common Cause Principle). The closest fact in this area that does provide explanatory information is this: that the barometer reading and the storm have a common cause. This reduces uncertainty about the causes responsible for the storm; it says that whatever the cause of the storm, it must be such that it also causes certain barometer readings. It excludes from the space of possible causes those that don't also cause certain barometer readings. I concede that this is explanatory information according to **ITE**, although obviously very limited.

An informational theory of explanation has been advocated by, e.g., Jackson and Pettit (1990), Lewis (1986), Railton (1981), and Skow (2014). Obviously there are differences between

[64] The informational version of the epistemic conception that Salmon (1984) attacked says that scientific explanations are "ways of increasing our information about phenomena of the sort we are trying to explain" (97), whereas **ITE** says that scientific explanations are ways of increasing our information about *the causes of* explananda. Salmon objected to the lack of causation in the former, not its reliance on the concept of information (1984, 101).

[65] I switch between talk of facts or events providing information and talk of representations of facts or events providing information.

these philosophers' theories. For my purposes, the relevant similarity is their emphasis on information. It is clear from their texts that they are not merely using the term “information” informally, but have in mind a notion like uncertainty- or possibility-reduction that generates the dilemma. For example, Lewis (1986: 217) writes that “to explain an event is to provide some information about its causal history”. Although he also writes that an explanation is “a proposition about the causal history of the explanandum event” (218), thereby dropping the term “information,” it is clear from his examples that a proposition about the causal history of the explanandum is one that provides information in the relevant sense (i.e. uncertainty- or possibility-reduction) about it. For example, explanatory information consists not only of positive information about what was in the causal history of the explanandum, but also negative information about what was not in the causal history of the explanandum (220). Lewis explicitly states that “the test” of whether something provides explanatory information “is that it suffices to rule out at least some hypotheses about the causal history of the explanandum” (221). This is precisely the sense of information as uncertainty- or possibility-reduction that generates the dilemma. Jackson and Pettit (1992) refer back to Lewis when they write, “We endorse the view that the job of explanation is to provide information on causal history” (13; Lewis is cited approvingly at p.12). The same idea is expressed when Skow (2014), explicitly following Lewis, writes that explanatory information “narrows down the list of possible causes (or possible causal histories) of the event being explained” (448) or “rule[s] out some hypotheses about what caused E [i.e. the explanandum]” (450). Finally, Railton (1981) also has uncertainty- or possibility-reduction in mind when he writes, “On the analysis given here, a proffered explanation supplies explanatory information (whether we recognize it as such or not) to the extent that it does in fact

(whether we know it or not) correctly answer questions about the relevant ideal text" (243). Here, though, it is not uncertainty about the causal history of the explanandum that is reduced in explanations, but uncertainty about the ideal explanatory text. This is especially clear when Railton writes, concerning the sense of information that he has in mind, that "the amount of information carried by a 'message' is proportional to the degree to which it reduces uncertainty. [...] Hence, information is a kind of *selection power* over possibilities" (244; original emphasis).

I have described an example of an informational theory of explanation and provided textual evidence that theories relying on the relevant notion of information have been prominent in the philosophy of explanation. Next I show how an informational theory is arguably implicit in core arguments of the mechanistic research program.

5.3. Information in the mechanistic research program

As far as I know, no prominent mechanists have endorsed an informational theory of mechanistic explanation[66]. However, one is arguably implicit in some of their core arguments. I illustrate this with Piccinini and Craver's (2011) argument that functional analyses are mechanism sketches and Zednik's (2011, 2015) argument that dynamical systems models can be mechanistic explanations.

5.3.1 Functional analyses as mechanism sketches

Piccinini and Craver (2011) argue that functional analyses are mechanism sketches. A functional analysis explains a system's ability or capacity in terms of the functional properties of the whole system or of its parts. Functional analysis is thought to proceed relatively autonomously of consideration of the structural components that realize the functional properties,

[66] Note that according to such a theory, two kinds of information could be explanatory: (causal) information about the mechanistic causal process that produces an explanandum result and (constitutive) information about the components, activities, and organization that constitute the mechanism that maintains, underlies, or produces the explanandum.

or play the functional roles, given the multiple realizability of such properties (Fodor 1968; Weiskopf 2011a,b). This provides prima facie reason for thinking that functional analyses are not mechanistic explanations of any kind. A mechanism sketch is an incomplete mechanistic explanation (Craver 2007). Mechanism sketches lack relevant details because they contain black boxes and filler terms.

Piccinini and Craver (2011) distinguish three types of functional analysis: task analysis, functional analysis by internal states, and boxology. I only briefly describe these, since Piccinini and Craver's argument is basically the same in each case. A task analysis decomposes a capacity into subcapacities and their organization (Cummins 1975). A functional analysis by internal states explains a capacity in terms of a system's internal states and their interaction (Fodor 1968). Boxology analyzes a system in terms of functional components or black boxes and their (often informational) interactions (Fodor 1968).

The reason Piccinini and Craver (2011) claim that all three kinds of functional analysis are mechanism sketches is that each puts constraints on the possible mechanisms that implement the functions (or subcapacities) identified in the analysis. Similarly, structure constrains function: not just any structural component can perform any function. For example, to perform the functions of belief and desire boxes, a mechanism(s) must be able to distinguish between those two types of representation and transform them in relevant ways (Piccinini and Craver 2011: 303). Although this is consistent with the multiple realizability of functions, it does put some limits on what could possibly implement belief and desire boxes.

When a functional analysis constrains mechanisms, it limits the space of possible mechanisms that could implement the identified functions.[67] But to reduce the space of possible

[67] It is possible that Piccinini and Craver could be working with a more robust notion of constraint that is

mechanisms just is to convey information about mechanisms. Therefore, Piccinini and Craver's argument is that functional analyses are mechanism sketches because they provide information about mechanisms. Piccinini and Craver appear to be relying on an informational theory of mechanistic explanation.[68]

5.3.2. Representational form as irrelevant to explanation

An informational theory is also implicit in, or at least implies the crucial premise of, Zednik's (2011, 2015) argument that dynamical systems models can provide mechanistic explanations.[69] Dynamical systems models employ the mathematical concepts of dynamical systems theory, such as differential or difference equations (Chemero 2009; Izhikevich 2007). These equations allow the modeling of the evolution of the target system's variables over time, which can be represented graphically as a trajectory through state space (or phase space). The state space of a system is a high dimensional space that represents all its possible states, i.e. all possible values of all the system's variables. Such graphical representations allow intuitive

---

inconsistent with an informational theory, but nothing in their paper suggests this.

[68] Piccinini and Craver's (2011) argument might not follow if Kaplan and Craver's (2011) model-to-mechanism-mapping (**3M**) requirement is true. According to **3M**, the variables in an explanatory model map onto specific structural components and activities of the explanandum phenomenon's mechanism, but you can constrain a mechanism without referring to it or its components and their activities, for example, by describing what the mechanism is *not* like. (Unless, of course, one has a liberal conception of properties according to which everything has an infinite number of negative properties.) Like mapping and reference, similarity is similarly too strong (Weisberg 2013) – accounts of explanation in those terms would not count as explanatory models that in fact are. This is because the information a model conveys – what can be learned from a model – and, so, its explanatory power, can outstrip what it explicitly represents. Yet, importantly, information about that on which an explanandum depends is all that is necessary to answer explanation-constituting w-questions. This contradicts **3M** even if it is consistent Piccinini and Craver's argument.
Also, that an informational theory is implicit in Piccinini and Craver (2011) could be especially problematic for Craver, according to whose interpretation of the ontic conception (2014), commitments about ontic structures are required to demarcate explanation from other scientific achievements. Part of Craver's motivation for this view is to avoid psychologistic accounts of explanation that require understanding. This may imply that whether or not a model explains is independent of the mental states of individual agents. If so, this seems to conflict with the fact that information is relative to an agent's background knowledge. My solution in Section 5 can resolve this apparent conflict.

[69] For arguments that dynamical models are not mechanistic explanations, see, e.g., Chemero 2009, Chemero and Silberstein 2011 and Stepp, Chemero, and Turvey 2011.

analysis of state space topology and reveal abstract, dynamical features such as attractors (states into which the system tends from surrounding states) (Izhikevich 2007).

Zednik's (2011) argument[70] distinguishes between mechanistic explanations, on the one hand, and the tools used for constructing and representing them, on the other. As a mathematical and conceptual framework, dynamical systems theory can be used to represent anything to which its concepts apply. If dynamical concepts can apply to the components, activities, and organization of mechanisms, then, according to Zednik, dynamical systems theory can provide mechanistic explanations. Zednik (2015) has extended this point, using evolutionary robotics and network science to show how new tools for mechanism description and discovery can go beyond the traditional mechanistic explanatory methods of decomposition and localization (Bechtel and Richardson 1993). What matters for an explanation to be mechanistic, according to Zednik, is not how it was constructed or its representational form – what matters is only that it represents a mechanism.

Although Zednik's argument may not directly imply an informational theory of mechanistic explanation, because representing a mechanism may be different than conveying information about it[71], it is consistent with and implied by an informational theory. This is because models of many different forms, constructed by many different methods, can provide information about mechanisms. Therefore, an informational theory of mechanistic explanation would imply that neither the form of a model nor the methods used to build it are relevant to whether it is a mechanistic explanation.

5.4. A dilemma for informational theories

[70] Unfortunately, going over Zednik's example would take too much space. I hope the tenor of the argument is clear without an example. For discussion, see Povich forthcoming b.

[71] See Dretske (1988) for an informational, teleological theory of representation.

Informational theories of explanation have often been proposed and are implicit in core arguments of the mechanistic research program. However, they seem to have some heretofore unseen counterintuitive consequences because information is most commonly thought to be relative to an agent's background knowledge. The relativity of information to background knowledge can be illustrated using Dretske's example of the shell game (1981, 78). In this example, there are four shells and a peanut is under one of them. Alice, but not Bob, knows that the peanut is not under shells 1 and 2. Alice and Bob then both turn over shell 3 and see that it is empty. This gives different information to Alice and Bob: Alice learns that the peanut is under shell 4, while Bob only learns that it is not under shell 3[72].[73]

Background knowledge, then, affects not only how much information is received, but *what* information is received (Dretske 1981, 81). In the most obvious and extreme case, an agent may not be able to extract any explanatory information from a model because she does not have the background knowledge to be able to interpret what the model says. Informational theories make it possible that a model could be explanatory to Alice, but not Bob, because it conveys information to Alice that it does not convey to Bob, due to differences in their background knowledge. It seems that the explanatoriness of a model can only be assessed relative to particular agents, contradicting commonsense judgments and traditional philosophical goals of the theory of explanation. In lay and scientific practice, we often assess the explanatoriness of a model *simpliciter*, not just relative to some particular agent. Similarly, one of the things a

[72] The difference in information conveyed to Alice and Bob can be precisely quantified (see Dretske 1981, 78). For Alice, 2 possibilities are reduced to 1, so she receives 1 bit of information; for Bob, 4 possibilities are reduced to 3, so he receives .42 bits of information.

[73] Julia Staffel (personal communication) objects that Alice and Bob receive the same information, but can draw different inferences because of their different background knowledge. I think this intuition is an artifact of this particular case, because the initial possibility space seems to come objectively pre-divided into four relevant possibilities. However, this is rare. Compare Dretske's discussion of the information generated by Edith's playing tennis (53) or the elimination of relevant possibilities (128).

philosophical theory of explanation should do is tell us when a model is an explanation *simpliciter*, not just when a model is an explanation relative to some particular agent. One desideratum of a theory of explanation should be to capture this "objectivity" of explanation.[74]

5.5. A solution to the dilemma

Informational theorists can avoid the above unwelcome consequences by relativizing to different sets of propositions and distinguishing scientific explanation *simpliciter* from explanation-to (i.e. explanation *to* a particular agent). For scientific explanation *simpliciter*, I propose to relativize information to "total scientific background knowledge" (TSBK), the current total store of propositions that are known the scientific community (cf. Kitcher [1989] on the explanatory store over *K*).

To determine the content of TSBK, I prefer Bird's (2010; 2014) account of scientific social knowledge. Bird conceives of social knowledge as performing social functions analogous to the functions of individual knowledge. So, for p to be socially known, p must be true, accessible to relevant members of the community (e.g., other scientists, but not necessarily the lay public), propositional in nature, the product of social mechanisms whose function promotes truth, and available as an input into social action or social cognitive structures (Bird 2010, 42–4). Unlike Bird, I remain agnostic as to whether there exists a single social agent that knows all the propositions that make up TSBK. Notice that while I have been speaking of information as relative to background knowledge, information is simply relative to a set of propositions – it need not be the case that that set of propositions is known by an agent.

When a model explains *to* an agent, I relativize to that particular agent's background knowledge. Explanation-to is where most pragmatic concerns are likely to arise. A model needs

[74] Of course, pragmatists such as van Fraassen (1980) may disagree. I will not here try to convince them otherwise.

to be presented in a particular way, tailored to the intended audience's presumed background knowledge, in order to maximize the probability that they will extract the appropriate information from it.

This distinction allows for it to be the case that a model is a scientific explanation *simpliciter*, but not an explanation *to* an individual agent, because their background knowledge does not allow them to extract the appropriate information from the model[75]. Metaphorically, we might say that even if a model does not inform an individual agent, it could inform the scientific community as a whole. If Bird (2010; 2014) is right, this would not just be metaphor – explanation *simpliciter* would be a form of explanation-to where the agent involved is a social agent.

Let me motivate why I am using Bird's account to determine the relevant set of propositions to which information is relativized in explanations *simpliciter* and not some other set of propositions, for example, the set of all true propositions or the set of propositions known at the "end of inquiry". To understand why I propose relativizing to TSBK, consider why it makes sense to relativize to background knowledge K^A, rather than, say, the set of all true propositions, in explanations to an agent A. It is only relative to A's own background knowledge that she is able to learn something from the information conveyed by a signal. On any particular occasion, A might not actually extract information from a model but she would be able to. An

[75] Eric Hochstein (personal communication) objects that a model can contain explanatory information for A without being an explanation to A, because A might not understand the model. To accommodate this intuition, I could distinguish between a model's being an explanation to A and a model's actually *explaining* to A. In the former, a model contains the relevant information relative to A, but A might not extract it. In the latter, A actually extracts the information. An analogous distinction seems possible with respect to scientific explanation *simpliciter*. For the scientific community to actually extract some information from a model means that that information becomes part of its knowledge, TSBK. Obviously, how a piece of information becomes part of TSBK is a controversial question. Bird (2010, 32) suggests publication as the relevant process by which a piece of knowledge becomes scientific knowledge.

analogous thought is at work in my TSBK proposal. It is helpful to think of the scientific community as something like a group agent with group knowledge, though, again, I do not want to commit to any robust notion of group agency. Although no single individual knows TSBK and although the scientific community might not actually extract information, and so learn something, from a model on any particular occasion, it would be able to, via the division of cognitive labor (Bird 2014). Bird's functional account of social knowledge ensures that social knowledge is useful to, and actually able to be used by, the scientific community. Explanatory *simpliciter* models inform the scientific community about, or increase the scientific community's knowledge of, causes (or mechanisms or the ideal explanatory text or whatever). If information were relativized to the set of all true propositions or the set of propositions known at the end of inquiry, we would be left with a conception of explanation that would never be of use to us, individually or collectively. Relativizing to TSBK captures just enough objectivity of explanation; any more would be inappropriate, given that models are interpreted representational structures used by us for our purposes.[76] I am willing to concede, therefore, that, for example, the Hodgkin-Huxley model of the action potential was not explanatory in the Middle Ages because no one, individually or collectively, had the ability to extract any information from it. We must either make this concession or, if we relativize to all true propositions or the end of inquiry, concede that there are explanatory models from which it is impossible for us, individually or collectively, to learn anything.

This solution can be expressed formally as follows. Think of the space of all possible causal histories of an explanandum phenomenon as a disjunction or set of those causal histories.

[76] Contessa (2007, 53 fn. 6) notes that one of the few agreements in the literature on scientific representation is that representation is a triadic relation between a representational vehicle, a target system, and a user base. It should be unsurprising that explanation is similarly implicitly triadic.

Since this set contains all possibilities, its prior probability is 1. When a model M conveys information about an explanandum phenomenon's causal history, it reduces the space of possible causal histories, thereby excluding some possibilities, i.e. decreasing the probability of a subset of initial possibilities to 0. This has the effect of increasing the probability of the subset of *remaining* possibilities to 1, since we now know that the actual causal history is in this subset. If we call a proper[77] subset of possible causal histories C_R, then a model M explains *simpliciter* when and only when $P(C_R|M\&TSBK)=1$, but <1 given TSBK alone (minus M; see below) (Dretske 1981). A model M explains to an agent A when and only when $P(C_R|M\&K^A)=1$, but <1 given K^A alone (minus M), where K^A is the agent's background knowledge.

I include the "minus M" condition because without it, a model would cease to provide information and, so, cease to be explanatory, once it becomes part of the current store of scientific knowledge or an individual agent's background knowledge. We want to say that such a model is still an explanation because it provides information relative to the relevant set of propositions, the model itself *excluded*.

If the probability of unity above is worrisome because it implies certainty, note that C_R is a *set* of causal histories (or mechanisms or propositions in the ideal explanatory text or whatever), not a particular causal history. The probability of unity here only implies that we can reduce the probability of some other possibilities (C_R's complement) to zero. If one is skeptical of this, we could instead require that M only probably exclude C_R's complement. This would reduce the probability of C_R's complement, but not to 0, increasing the probability of C_R, but not to 1 (see Scarantino 2015). Then, a model M explains *simpliciter* when and only when $P(C_R|M\&TSBK)>P(C_R|TSBK–M)$. A model M explains to an agent A when and only when $P(C_R|$

[77] C_R must be a proper subset or no information is conveyed.

$M\&K^A)>P(C_R|K^A–M)$.

5.6. Conclusion

Informational theories have an impressive history in the philosophy of explanation (e.g., Jackson and Pettit 1990; Lewis 1986; Railton 1981; Skow 2014). I have shown how an informational theory is also implicit in core arguments of the mechanistic research program. An informational theory is implied by Piccinini and Craver's (2011) arguments that functional analyses are mechanism sketches; an informational theory is at least consistent with and implies the central premise in Zednik's (2011, 2015) argument that dynamical systems models can be mechanistic explanations.

Informational theories, however, seem to have counterintuitive consequences, for they seem to imply that assessments of explanatoriness can only be made relative to individual agents. This conflicts with lay and scientific commonsense judgments about explanation and traditional goals of the theory of explanation. By distinguishing scientific explanation *simpliciter* from explanation-to and relativizing the former to TSBK, these counterintuitive consequences are avoided.

Chapter 6. Social Knowledge and Supervenience Revisited

Bird's (2010; 2014) account of social knowledge (**SK**) denies that scientific social knowledge supervenes solely on the mental states of individuals. Lackey (2014) objects that **SK** cannot accommodate 1) a knowledge-action principle and 2) the role of group defeaters. I argue that Lackey's knowledge-action principle is ambiguous. On one disambiguation, it is false; on the other, it is true but poses no threat to **SK**. Regarding group defeaters, I argue that there are at least two options available to the defender of **SK**, both taken from literature on individual defeaters and applied to group defeaters. Finally, I argue that Lackey's description of the case of Dr. N. – as a case in which the scientific community does not know but is merely in a position to know – is mistaken. It assumes that Dr. N.'s publication is not scientific knowledge. An analogy to the individual case shows that it is plausible that the scientific community is not merely in a position to know, although its members are. This leaves intact a conception of social knowledge on which it does not supervene on the mental states of individuals.

6.1. Introduction

Bird (2010; 2014) has defended a radically extended account of scientific social knowledge. It is radically extended in the sense that it denies that group knowledge supervenes on the mental states of individuals. Let us, following Lackey (2014), call this account **SK**. Lackey (2014) objects that **SK** cannot accommodate 1) a knowledge-action principle (when combined with another plausible principle linking individual and group action) and 2) the role of group defeaters. I argue that Lackey's knowledge-action principle is ambiguous. On one disambiguation, it is false; on the other, it is true but poses no threat to **SK**. Regarding group defeaters, I argue that there are at least two options available to the defender of **SK**, both taken

from literature on individual defeaters and applied to group defeaters. Finally, I argue that Lackey's description of the case of Dr. N – as a case in which the scientific community does not know but is merely in a position to know – is mistaken. Lackey's analogy to the individual case can be redescribed to show that it is plausible that the scientific community is not merely in a position to know, although its members are.

In Section 6.2, I present more completely **SK**; this Section also spells out the central example of non-supervenience under dispute – the case of Dr. N**.** In Sections 6.3 and 6.4, I present and respond to Lackey's action and defeater objections, respectively. In Section 6.5, I show how Lackey's distinction between knowing and merely being in a position to know, while useful, is misapplied to the case of Dr. N.

6.2. Scientific Social Knowledge

When does a social epistemic subject such as the scientific community know?[78] According to **SK** (Bird 2010; 2014), a social epistemic subject S knows that p if and only if p is true, the information that p is accessible to relevant members of S (e.g., other scientists, but not necessarily the lay public), and the following three conditions hold[79]:

> (i) [S has] characteristic outputs that are propositional in nature (*propositionality*);
> (ii) [S has] characteristic mechanisms whose function is to ensure or promote the chances that the outputs in (i) are true (*truth-filtering*);
> (iii) the outputs in (i) are the inputs for (a) social actions or for (b) social cognitive structures (including the very same structure [the structure that produces the output]) (*function of outputs*). (Bird 2010, 42–4)

The three conditions are supposed to be social analogues of the functions of and conditions required for individual knowledge. I will not defend them here. My focus instead will be on

[78] My focus will not be on the ontology of social epistemic subjects. See Bird (2010; 2014) for an argument that the division of cognitive labor in science enforces an organic interdependence that constitutes it as a social subject.

[79] Bird does not give necessary and sufficient conditions for social knowledge. This biconditional is Lackey's (2014) reconstruction of Bird, which I endorse here, both as a reconstruction of Bird and as an account of social knowledge.

responding to Lackey's (2014) objections, which do not target these conditions.

Bird argues that **SK** implies that scientific knowledge does not supervene on the mental states of individuals. This is illustrated by Bird's cases of Drs. Q. and N. These cases are identical with respect to the mental states of individuals, but differ with respect to social knowledge. Dr. Q. is a reclusive scientist who dies with her results tucked away from the scientific community. Even though Dr. Q. knew that p, the scientific community did not, in virtue of her lack of publication. This is because publication is what makes a proposition accessible to the scientific community (Bird 2010, 32). Dr. N., on the other hand, is a scientist whose results, though published in a reputable journal, everyone forgets once she dies. Bird (2010, 32) claims that, in virtue of publication and, so, accessibility, the scientific community knows Dr. N.'s conclusion that d, even though no individual scientist knows that d.[80]

The cases of Drs. Q. and N. suggest that the set of propositions known by the scientific community is independent of what individuals know. Individual knowledge is insufficient, as shown by the case of Dr. Q., and unnecessary, as shown by the case of Dr. N., for social knowledge. I will defend Bird's treatment of the case of Dr. N. from objections in the next two sections.

6.3. Lackey's Objections: Action

Lackey (2014) argues that the following two principles conflict with **SK**.[81]

> (**KAP**) Knowledge/Action Principle: S knows that p only if S is epistemically rational to act as if p or, equivalently, S is epistemically rational to act as if p if S knows that p. (287)[82]

[80] As Bird (2010) describes the case, eventually Dr. N.'s publication is rediscovered, but its rediscovery is unnecessary for social knowledge. All that is required is its rediscoverability (i.e. accessibility).

[81] If Carter (2015) is right, then Lackey's objections extend beyond **SK** to some mainstream views on group knowledge. Thanks to an anonymous reviewer for bringing this to my attention.

[82] One might respond to Lackey's objection by denying **KAP** (Brown 2008; Lackey 2010), though I will not do so here. Thanks to an anonymous reviewer for pointing this out.

> (**GMAP**) Group/Member Action Principle: For every group, G, and act, a, G performs a only if at least one member of G performs some act or other that causally contributes to a. (286)

According to Lackey (2014), these two principles are inconsistent with **SK**, as shown by applying them to the above-mentioned case of Dr. N., whose results, though published in a reputable journal, everyone forgets once she dies. In this case, Bird and Lackey agree that **SK** implies that the scientific community knows that d even though no living individual does. Applying **GMAP**, if the scientific community were to act, it would be through members unaware of d. Lackey (2014) argues that it is therefore epistemically irrational for the scientific community to act as if d, for example, by asserting[83] d or approving drugs that depend on d[84]. But, given **KAP**, the epistemic irrationality of acting as if d shows that the scientific community does not know d. Hence, **GMAP** and **KAP** show that the scientific community does not know that d; **SK** is false.

In response, I argue that **KAP** is ambiguous between two different sense of "act as if". I distinguish "acting on" knowledge that p from merely "acting in accordance with" knowledge that p. On the latter disambiguation, **KAP** is false of individual and social agents. On the former disambiguation, **KAP** is true, but does not show that **SK** is false.

The disambiguation has to do with knowledge's role in action production. Consider Hawthorne and Stanley's reasoning when defending their version of **KAP**: "When someone acts on a belief that does not amount to knowledge, she violates the norm, and hence is subject to criticism" (Hawthorne and Stanley 2008, 577). The phrase "acts on" seems to me to imply that

[83] Lackey uses assertion as an example, but note that knowledge-action principles are usually distinguished from knowledge norms of assertion. Thanks to Julia Staffel for bringing this to my attention.

[84] Note that this says that it is irrational to act as if d while being unaware of d. I will assume this is true. It does not say – what is likely false – that it is irrational to act as if d while being unaware of one's knowledge that d. See Williamson's (2000) anti-luminosity arguments.

the belief plays some causal role in producing the subject's action[85]. This is the plausible disambiguation of "act as if". The less plausible disambiguation treats acting as if p as merely acting in accordance with p, where it could be sheer luck that one's action so accords. In merely acting in accordance with p, one's knowledge that p plays no role in the action.[86] Compare two cases: 1) agent A knows that p and merely acts in accordance with p, for example, asserting that p though her knowledge plays no causal role in the production of her assertion, and 2) agent A is in an exactly similar situation, except her knowledge that p does play a causal role in the production of her assertion. A's action in the former case is not epistemically rational, but her action in the latter case is. This suggests that the proper formulation of **KAP** is something like: If S knows that p, then S is epistemically rational to act *on* p (not merely in accordance with p). **SK** is consistent with this.

Now we can apply this distinction to the case of a social epistemic subject. The latter, epistemically irrational scenario above is analogous to Lackey's description of the case under dispute. In Lackey's example, the reason the scientific community is irrational for asserting d or approving drugs that depend on d is that, in my terms, the scientific community is merely acting in accordance with d, rather than acting on d. In Lackey's example, the scientific community's knowledge that d plays no role in the production of its actions. I agree that this is epistemically irrational, but it is also irrational in the analogous case of individual knowledge, as shown above. In both the social and individual cases, the irrationality stems from the fact that acting in mere accordance with p would be a matter of epistemic luck (cf. Lackey 289).

I claim that in the case under dispute the scientific community is merely acting in

[85] This role is more explicit in other formulations of knowledge-action principles that stress knowledge's role in practical reasoning (e.g., Hawthorne and Stanley 2008), where knowledge must interact with other mental states in rational and causal processes. Lackey's formulation of **KAP** downplays this.

[86] Cf. Turri (2011), who distinguishes having a reason and believing for a reason.

accordance with d, rather than acting on d, because knowledge that d is not playing a role in producing social action. What does it mean for knowledge that d to play a role in producing the scientific community's action? Let us briefly look at the individual case first. For knowledge that d to play a role in producing an agent A's action, A does not have to be aware of her knowledge that d. What is required is that her knowledge that p play an appropriate (i.e. non-deviant)[87] causal role in producing her action. Presumably, what this entails is something like that her brain and motor system are wired up such that the brain state on which her knowledge supervenes is non-deviantly causally relevant to her motor output. So, to apply this to the social case, we must have some account of the supervenience base of scientific social knowledge, so that we can see how it is causally relevant to social action. In social cases, the supervenience base of knowledge can include the mental states of individuals, but it also includes social institutions and objects like articles, laboratories, and the internet (Bird 2010; 2014). For knowledge that d to play a role in producing the community's action, then, is for the community to make use of these knowledge-supervening social structures in its decision-making (see condition iii of **SK** above). Since one of the relevant structures – perhaps the only structure – on which the social knowledge that d supervenes is Dr. N.'s publication, the only way for the community to act *on* d is for Dr. N.'s publication to play some role in the community's decision-making. This is exactly what is ruled out in Lackey's description of the case; Dr. N.'s publication plays no role in the community's social actions. The way Lackey describes the case of Dr. N. is therefore not a case of the scientific community acting *on* its knowledge that d, but merely acting in accordance with it. I therefore conclude that **SK** is consistent with **KAP** properly construed to

[87] I set aside worries about causal deviance, i.e. whether and how her knowledge must cause her action *in the right way*. Whatever the right account is in the individual case, it can likely be extended to the social case (e.g. Turri 2011).

include a causal role for knowledge in action production; it would be rational for the scientific community to act *on* d, even if no individual is aware that d.

It would be rational for the scientific community to act on d, even if no individual is aware that d, but how could it? That is, how could the social knowledge that d play a non-deviant role in the community's social action without any of its members being aware that d? One of Bird's (2010) examples of scientific knowledge without awareness could be contrived to allow for this, especially if reliable automation is involved (35). For example, consider Lackey's description of the case of Dr. N. again, where no one is aware of Dr. N.'s finding that d though the community makes d-relevant decisions. Now, let us add that the scientific community has created a reliable, automated system that checks past results for certain truth-conducive properties like statistical significance, statistical power, effect size, etc. If the automated system detects a result that passes the community's standards, it sends the result to another automated system, reliable and trusted by the community, that uses the result to help the community make social decisions. The first system detects Dr. N.'s results and sends them to the next system for processing in social action. Here it is plausible that the knowledge that d plays a (non-deviant) causal role in the production of social action without any member being aware that d.[88] As Bird (2010, 35) notes, cases like this are becoming increasingly likely as the scientific division of labor becomes ever more divided and automated.

6.4. Lackey's Objections: Defeaters

Lackey's (2014) second objection to the case of Dr. N. has to do with defeaters. She distinguishes psychological defeaters and normative defeaters[89] (2014, 292). A psychological

[88] Carl Craver (personal communication) helpfully suggests that this case might be analogous to individual cases of self-deception.

[89] She does not mention externalistic defeaters. I return to this below.

defeater is a mental state of S's that rebuts S's belief that p or undercuts her justification for believing that p. A normative defeater is a psychological defeater that S should have, given available evidence. The objection begins with an addition to the case of Dr. N. The addition is that the vast majority of individual members of the scientific community, via some undescribed process, come to believe that not-d. Lackey then argues that this implies that the scientific community itself believes that not-d and its belief that not-d defeats its putative knowledge that d.[90]

Lackey (293) draws two conclusion from this. First, social knowledge may be a lot less common than we would like. For many scientific propositions, there could be a number of scientists who hold conflicting beliefs that could function as psychological defeaters. Second, if the belief that not-d defeats the knowledge that d, this implies an arbitrary asymmetry: group beliefs cannot justify other group beliefs, but they can defeat other group beliefs, which, she claims, is arbitrary without a story about why.

However, it is not clear why Lackey believes that in the case described, it is the belief that not-d that must defeat the knowledge that d. She does not, for example, claim that the belief that not-d is more justified than the belief that d. She does not explain why everyone believes that not-d. It is open for the proponent of **SK** to claim that the knowledge that d defeats the belief that not-d. After all, the knowledge that d is, we are supposing, true, reliably formed, and easily accessible to all the relevant scientists, while the belief that not-d is a widespread falsehood. There are accounts of defeat where easily accessible but unpossessed evidence can be a defeater (e.g., Harman 1973) or where the availability of an alternative reliable process that, had it been

[90] Like Lackey, I will ignore the distinction between *prima facie* and *ultima facie* justification. This should not affect the points that follow.

used in addition to the actual process, would have led to a different belief can be a defeater (e.g., Goldman 1979).

Goldman (1979) explicitly excludes gathering new evidence from alternative reliable processes. This might seem to preclude the application of Goldman's account of defeat to this case. However, note that, according to **SK**, accessing Dr. N.'s result that d is not the scientific community gathering new evidence; it is accessing knowledge it already has, though its members might not have. The social situation is somewhat akin to introspection, an individual accessing a memory of previously acquired evidence, which Goldman endorses as an alternative reliable process. Of course, this process of "social introspection," accessing previous scientific results, must itself be a reliable process for the reliabilist account of defeat to work in this case. However, I think it is plausible that it is, given the kinds of truth-filtering mechanism that ensure condition ii of **SK** (Bird 2010, 43–4). There is therefore a reliabilist story that is not ad hoc about why the knowledge that d defeats the belief that not-d.[91] This is the first strategy for the defender of **SK**.

Suppose we adjust the case of Dr. N. so that all the members of the scientific community possess misleading evidence supporting their beliefs that not-d. Would not these individuals' beliefs also constitute a defeater of the community's knowledge that d?[92] I do not think that claiming that there is mutual defeat *per se* in this case is a problem for **SK**. Recall that Lackey presents her example to suggest that there is an arbitrary asymmetry for the defender of **SK**:

[91] I do not wish to claim that the existence of alternative reliable processes accounts for *every* case of group defeat. There are well-known problems of method individuation (i.e. the generality problem) for reliabilism. See, for example, Baker–Hytch and Benton (2015), Beddor (2015), and Lasonen–Aarnio (2010) for arguments that this will not work in every case of defeat. I only claim that it is plausible that the defender of **SK** can use the reliabilist story to respond to Lackey's example (see also Goldman 2014). See Grundmann (2009) for an argument that reliabilism can accommodate defeaters.

[92] I thank two anonymous reviewers for pressing this objection.

group beliefs cannot justify other group beliefs, but they can defeat other group beliefs. Since that is the conclusion I am trying to avoid, I think I can consistently accept that there is mutual defeat in this example, as long as such defeat is not conceived internalistically as effected by group (or individual) beliefs. Rather, the availability of Dr. N.'s publication serves as a defeater for the group belief that not-d and the availability of other publications misleadingly supporting not-d may serve as a defeater for the group belief that d. However, I am still open to Lackey's objection that social knowledge may be less common than we would like. Maybe so, but it seems to me that, given the extent of inconsistent evidence and controversy in scientific practice, this may be a result of any plausible account of group knowledge in science. It is also not clear that Lackey's (2016) positive account of justified group belief fares better than Bird's when it comes to the *amount* of scientific social knowledge it allows. In particular, the prevalence of disagreement in science seems to suggest that it will be rare that "a significant percentage of the operative members" of the scientific community justifiedly believe that p and that full disclosure of and rational deliberation about all the relevant evidence for p "would not result in further evidence that when added to the bases of [the scientific community]'s members' beliefs that p, yields a total belief set that fails to make sufficiently probable that p" (381). However, an in-depth discussion of Lackey's positive account is beyond the scope of this paper.

The second response to Lackey is to deny the existence of group defeat, not because there is something wrong with *group* defeat, but because there is something wrong with defeat itself. Lackey says that if there is no defeater in this case, then this would remove groups "from the realm of the rational altogether" because "rebutting defeaters are precisely what rule out this combination of states from being epistemically permissible" (294). However, the defender of **SK**

can account for the irrationality of this case by appeal to a kind of knowledge norm of theoretical reasoning similar to **KAP** above (Baker–Hytch and Benton 2015; see also Lasonen–Aarnio, M. 2010).[93] The norm would be something like Baker–Hytch and Benton's knowledge norm of belief (Williamson 2000):

> (**KNB**) One must: not believe that p if one does not know that p. (Baker–Hytch and Benton 2015, 56)

The upshot of this strategy is that the defender of **SK** need not say that the scientific community's knowledge that d defeats its belief that not-d (or vice versa). The intuition that something has gone wrong in this case is better explained by the community's failure to adhere to **KNB**. The community possesses evidence that its belief that not-d is false, which means that it possesses evidence that its belief that not-d is not knowledge. In such a situation, **KNB** rules that the community must not believe that not-d. It is **KNB**, not rebutting defeaters, that rules out this combination of states from being epistemically permissible.

Indeed, although the following claim is not necessary for a response to Lackey, the scientific community appears to adhere to something like **KNB**. (Whether or not the scientific community actually adheres to this norm is different from whether or not it applies.) However, the scientific community seems consistently to engage in behavior that follows the norm, for example, by issuing retractions of published results that come under dispute and by not letting disputed beliefs guide its social actions.

6.5. Is the Scientific Community Merely in a Position to Know?

Having argued against the claim that the scientific community knows that d in the case of

[93] Bird would likely approve of this response to Lackey since he explicitly compares his functionalist approach to knowledge-first epistemology, writing that, "According to my view, the function of the cognitive faculties is just that, to provide a link between the subject and the relevant facts so that they may be used as the inputs (reasons) in practical and theoretical reasoning" (2010, 42).

Dr. N., Lackey argues that the case is more accurately described as one in which the community is merely in a position to know. According to Lackey, the distinction between knowing and merely being in a position to know "is grounded, at least in part, in the difference between information that has been accessed and information that is merely accessible" (2014, 294). She gives the following example of an individual merely being in a position to know and claims that it is analogous to the community in the case of Dr. N. On my desk there is an unopened letter my friend's confession to a crime. I know nothing of the crime beforehand. In this case, I do not know that my friend committed the crime, though I am in a position to know by opening it and reading. The scientific community, Lackey claims, is in a similar position with respect to Dr. N.'s publication.

The proponent of **SK** is likely to find Lackey's analogy question-begging, for it assumes that the information in the letter – the analog of Dr. N.'s publication – is not known. But whether Dr. N.'s publication is known by the community is what is under dispute. Let me end by offering a reconceptualization of this analogy in light of my arguments above. The case of the unopened latter is not analogous to the case of Dr. N. An individual analog to the case of Dr. N. would be a case where the information in the unopened letter were part of the individual's knowledge. The individual is not merely in a position to know that p, but knows that p, although he or she might not be aware of that fact. The accessibility of Dr. N.'s publication makes it knowledge for the scientific community. It is analogous to an individual's accessible but currently unaccessed memory that p.[94] The individual members of the scientific community are merely in a position to know that d, but the scientific community itself already does know that d.

[94] A reviewer notes another analogy to individual memory: just as an individual will scan her memory for relevant evidence before embarking on important action, so will the scientific community scan what it knows (its published articles).

6.6. Conclusion

The cases of Drs. Q. and N. are intended by Bird (2010; 2014) to demonstrate the non-supervenience of group knowledge on individual knowledge. Lackey (2014) argues that this non-supervenience cannot accommodate **KAP** (plus **GMAP**) and the role of group defeaters. I argued that "act as if" in **KAP** is ambiguous between "acting on" the knowledge that p and merely "acting in accordance with" the knowledge that p. On the latter disambiguation, **KAP** is false for individuals and groups. On the former disambiguation, **KAP** is true, but does not show that **SK** is false. Regarding group defeaters, I argued that the defender of **SK** could either claim that there is a reliabilist defeater in Lackey's example or that, appealing to **KNB**, there are no defeaters. If this is right, then non-supervenience remains a viable position in the metaphysics of social knowledge.

Chapter 7. The Directionality of Distinctively Mathematical Explanations

with Carl F. Craver

In “What Makes a Scientific Explanation Distinctively Mathematical?” (2013b), Lange uses several compelling examples to argue that certain explanations for natural phenomena appeal primarily to mathematical, rather than natural, facts. In such explanations, the core explanatory facts are modally stronger than facts about causation, regularity, and other natural relations. We show that Lange's account of distinctively mathematical explanation is flawed in that it fails to account for the implicit directionality in each of his examples. This inadequacy is remediable in each case by appeal to ontic facts that account for why the explanation is acceptable in one direction and unacceptable in the other direction. The mathematics involved in these examples cannot play this crucial normative role. While Lange's examples fail to demonstrate the existence of distinctively mathematical explanations, they help to emphasize that many superficially natural scientific explanations rely for their explanatory force on relations of stronger-than-natural necessity. These are not opposing kinds of scientific explanations; they are different aspects of scientific explanation.

7.1. Introduction.

In “What Makes a Scientific Explanation Distinctively Mathematical?” (2013b; 2013a), Lange uses several compelling examples to argue that certain natural phenomena are best explained by appeal to mathematical, rather than natural, facts. In distinctively mathematical explanations, the core explanatory facts are modally stronger than facts about, e.g., statistical relevance, causation, or natural law. A distinctively mathematical explanation might describe causes, Lange allows, but its explanatory force derives ultimately from appeal to facts that are

"more necessary" than causal laws. Lange advances this thesis to argue for the importance of a purely modal view of explanation (a view that emphasizes necessities, possibilities, and impossibilities, showing that an event had to or could not have happened) in contrast to the widely discussed ontic view (a view that associates explanation with describing the relevant natural facts, e.g., about how the event was caused or how its underlying mechanisms work).[95]

Lange operates with a narrower understanding of the ontic conception. He describes it as the view that all explanations are causal. He cites Salmon, who claimed that, "To give scientific explanations is to show how events and statistical regularities fit into the causal structure of the world" (Salmon 1984)[96] and "To understand why certain things happen, we need to see how they are produced by these mechanisms [processes, interactions, laws]" (Salmon 1984). He also cites Lewis ("Here is my main thesis: to explain an event is to provide some information about its causal history"; 1986) and Sober ("The explanation of an event describes the 'causal structure' in which it is embedded"; 1984).[97] In contrast to Lange, we adopt a more inclusive understanding of the ontic that embraces any natural regularity (Salmon 1989; Craver 2014; Povich forthcoming a), e.g., statistical relevance (Salmon 1977), natural laws (Hempel 1965), or contingent

[95] There is a growing body of literature on mathematical explanation (Baker 2005; Baker and Colyvan 2011; Huneman 2010; Pincock 2011). We focus on Lange because his examples have become canonical and because his commitments are so explicitly formulated. We suspect that the directionality problem will arise in these other papers as well, but these authors are mostly concerned with indispensability and the ontology of mathematics, a topic that we (like Lange) hope to sidestep to focus on explanation alone. See Craver (forthcoming) for a discussion of directionality problems in network explanation. Andersen's (forthcoming) response to Lange is complementary to ours, fleshing out a point about explananda at which we only gesture in the conclusion. Our main focus is directionality.

[96] See the passages quoted in Povich (forthcoming) for evidence that Salmon did not think the ontic conception was strictly causal. As we note, Lange's conception of the ontic conception is narrower than one might allow. The primary aim of the ontic conception is to insist that whether X explains Y is an objective matter of (natural) fact.

[97] One can believe that mechanistic explanation is important without believing that all explanations are causal or mechanical. We show why $C = 2\pi r$ without describing mechanisms. We explain why Obama can sign treaties without describing causes. Explanations in epistemology, logic, and metaphysics often work without describing causes. The question here is not whether one should be a pluralist about explanation but about whether Lange's account of distinctively mathematical explanation is complete and whether his contrast with the ontic conception is substantiated by his examples.

compositional relations might also figure fundamentally in explanation. This point will become crucial below, given that the ontic relations that explain the directionality of some explanations are not specifically causal relations; but they are ontic in this wider sense.[98] Lange's arguments should, however, work equally well against this broader understanding of the ontic conception, given that he uses the examples to show that some explanations of natural facts depend fundamentally on relations of necessity that are stronger than mere natural necessity.

We argue that Lange's account of distinctively mathematical explanation is flawed. Specifically, it fails to account for the directionality implicit in his examples of distinctively mathematical explanation. This failure threatens Lange's argument because it shows that his examples do not, in fact, derive their explanatory force from mathematical relations alone (independent of ontic considerations). The inadequacy is in each case easily remediable by appeal to ontic facts that account for why the explanation is acceptable in one direction and unacceptable in the other. That is, Lange's exemplars of distinctively mathematical explanation appear to require for their adequacy appeal to natural, ontic facts about, e.g., causation, constitution, and regularity. More positively, we suggest that all mechanistic explanations are constrained, and so explained, by both ontic and modal facts. Rather than seeing an opposition between distinctively mathematical explanations and causal (or more broadly ontic) explanations, Lange's examples, as we reinterpret them, direct us to understand how these distinct aspects of explanation, these distinct sources of explanatory power, intermingle and interact with one another in most scientific explanations.

[98] For purposes of focus, we leave aside the question of whether the existence of distinctively mathematical explanations in fact commits one to the denial of the ontic conception or even to the idea that there is a modal form of explanation independent of ontic considerations. The fact that mathematics is important to explanation doesn't necessarily commit one to the idea that the modal conception has a role to play independently of ontic considerations absent further commitments about the relationship between mathematics and ontology. Like Lange, we remain silent on the ontology of mathematics (492).

7.2. Lange's Account of Distinctively Mathematical Explanation.

Lange's goal is to show “how distinctively mathematical explanations work” by revealing the “source of their explanatory power” (486). He accepts as a basic constraint on his account that it should “fit scientific practice,” that is, that it should judge as “explanatory only hypotheses that would (if true) constitute genuine scientific explanations” (486). In short, the account should not contradict too many scientific common-sense judgments about whether an explanation is good or bad. Lange's goal and his guiding constraint are conceptually related: to identify the source of an explanation's power requires identifying the key features that sort acceptable explanations from unacceptable explanations of that type. In causal explanations, for example, much of the explanatory power comes from knowledge of the causal relations among components in a mechanism. Bad causal explanations of this kind fail when they misrepresent the relevant causal structure (in ways that matter). In mathematical explanations, on Lange's view, the explanatory force comes from mathematical relations that are 'more necessary' than mere causal or correlational regularities.

Given this set-up, Lange's account of the explanatory force of distinctively mathematical explanations can be undermined by examples that fit Lange's account but that would be rejected as bad explanations as a matter of scientific common-sense. The account would fail to identify fully the explanatory force in such explanations and so would fail to account for the norms governing such explanations.

Lange does not address the canonical form of mathematical explanations. However, his examples are readily reconstructed as arguments in which a description of an explanandum phenomenon follows from an empirical premise (EP) describing the relevant natural facts, and a

mathematical premise (MP) describing one or more more-than-merely-naturally-necessary facts. To begin with Lange's simplest example:

> **Strawberries:** Why can't Mary divide her strawberries among her three kids?[99] Because she has 23 strawberries, and 23 is not divisible by three.

This explanation can be reconstructed as an argument:

> 1. Mary has 23 strawberries (EP)
> 2. 23 is indivisible by 3 (MP)
> C. Mary can't divide the strawberries equally among her three kids.[100]

We would have to tighten the bolts to make the argument valid (e.g., no cutting of strawberries is allowed), but the general idea is clear enough. The empirical premise works by describing the natural features of a system. They specify, for example, the relevant magnitudes (Mary starts with 23 strawberries), and the causal or otherwise relevant dependencies among them. All mathematical explanations of natural phenomena require at least some empirical premises to show how the mathematics will be applied and to specify the natural (empirically discovered) constraints under which the mathematical premises do their work. The question is whether those mathematical premises are supplying the bulk of the 'force' of the explanation, as appears to be the case in Strawberries.[101]

Lange's other examples can similarly be reconstructed as arguments mixing empirical and mathematical premises.

---

[99] Or “Why didn't she on some particular occasion?” or “Why didn't or couldn't anyone ever?” Lange intends all these explananda to be explained by the same explanans; a similar multiplicity of explananda can be generated for the examples below.

[100] This example is reconstructed as a sketch of a deductive argument, but distinctively mathematical explanations might be inductive. For example, one might explain why fair dice will most likely not roll a string of ten consecutive double-sixes on mathematical grounds, using logical probability and some math.

[101] Lange might object to the inclusion of the empirical premise in this formulation. Instead, he might treat the empirical premise as a presupposition of the why question: “Why can't Mom divide her 23 strawberries among her three kids?” Answer: “Because 23 is indivisible by 3.” In what follows, all of our examples can be so translated without affecting the principled incompleteness in the cases, but this reformulation comes at considerable cost to the clarity with which the incompleteness can be displayed (see Section 7.4).

Trefoil Knot: Why can't Terry untie his shoes? Because Terry has a trefoil knot in his shoelace (EP). The trefoil knot is not isotopic to the unknot in three dimensions (EP), and only knots isotopic to the unknot in three dimensions can be untied (MP) (489).

Königsberg: Why did Marta fail to walk a path through Königsberg in 1735, crossing each of its bridges exactly once (an Eulerian walk)? Because, that year, Königsberg's bridges formed a connected network with four nodes (landmasses); three nodes had three edges (bridges); one had five (EP). But only networks that contain either zero or two nodes with an odd number of edges permit an Eulerian walk (MP) (489).

Chopsticks: Why is it likely that more tossed chopsticks will be oriented horizontally rather than vertically? Because the they were tossed randomly (EP) and there are more ways for a chopstick to be horizontal than to be vertical (MP). If we focus on the sphere produced by rotating the chopstick through three dimensions, a chopstick can be horizontal anywhere near the equator; it is vertical only near the poles (490).

Cicadas: Why do cicadas with prime life-cycle periods tend to suffer less from predation by predators with periodic life cycles than do cicadas with composite periods? Because it minimizes predation to have a life cycle that intersects only infrequently with that of your periodic predators (EP) and because prime periods minimize the frequency of intersection (MP) (498).

Honeycombs: Why do honeybees use at least the amount of wax they would use to divide their combs into hexagons of equal area? Because honeybees divide their combs (which are planar regions with dividing walls of negligible thickness) into regions of equal area (EP) and a hexagonal grid uses the least total perimeter in dividing a planar region into regions of equal area (MP) (499).[102]

Pendulum: Why does Patty's pendulum have at least four equilibrium configurations? Because Patty's pendulum is a double pendulum (EP) and any double pendulum's configuration space is a torus with at least four stationary points (MP) (501).

Central to Lange's broader purposes is the claim that these distinctively mathematical explanations gain their explanatory force from non-causal, and more broadly, non-ontic sources: i.e., stronger-than-naturally-necessary relations. Explanatory priority flows downward from the more necessary to the less necessary:

In my view, the order of causal priority is not responsible for the order of explanatory

[102] Lange "narrows" the explananda in these last two cases. Note that the explananda are *not*, respectively, that cicadas have prime periods and that honeycombs are hexagonal. Those explananda have causal (etiological and constitutive) explanations. The narrower explanations, Lange argues, have distinctively mathematical explanations.

> priority in distinctively mathematical explanations in science. Rather, the facts doing the explaining are eligible to explain by virtue of being modally more necessary even than ordinary causal laws (as both mathematical facts and Newton's second law are) or being understood in the why question's context as constitutive of the physical task or arrangement at issue. (506)

For Lange, distinctively mathematical explanations gain their explanatory force from the fact that they rely fundamentally on mathematical relations that are more necessary than are relations of causation and natural law. The norms by which good mathematical explanations are sorted from bad mathematical explanations would, according to this account, turn on the relevant mathematics and facts about how that mathematics is being applied. In the following section we argue that Lange's analysis is inadequate.

7.3. The Inadequacy of Lange's Model.

Lange's account currently leaves unspecified a crucial feature for sorting the mathematical arguments that have explanatory power from those that fail as explanations. Our argument for this thesis is inspired by Bromberger's example of the flagpole and the shadow (1966). At least according to scientific common-sense, one can explain the shadow's length by appealing to the flagpole's height, the sun's angle of elevation, and the natural fact that light propagates in straight lines. One cannot, in accord with scientific common-sense, explain the flagpole's height in terms of its shadow's length, the angle of the sun's elevation, and the natural fact that light propagates in straight lines (in non-intentional contexts; cf. van Fraassen 1980, 132–4). In complete accordance with the norms of the once-received, covering-law model of explanation (Hempel 1965), one can write a deductive argument relating law statements and true descriptions of 'initial' conditions to either conclusion. For Bromberger (and Salmon 1984), the example demonstrates an *asymmetry* in *natural* explanations that the covering-law model could

not accommodate. The covering-law model is thereby shown to be an inadequate account of the norms of scientific explanation.

More generally, the example demonstrates that at least some (and in fact, many) explanations have a preferred direction. Salmon, for example, used this example (among others) to argue that scientific explanations work by tracing the antecedent causal structure of an event: Light leaves the sun, passes the flagpole, and lands on the ground. Causation enforces this temporal direction. No such causal sequence proceeds from the shadow to the height of the flagpole (outside intentional contexts). Considerations of just this sort underlie both Lewis' and Sober's emphasis on causation as the fundament of scientific explanation. In what follows, we emphasize the *directionality* of explanations, not their asymmetry. It does not matter for our purposes whether all the same statements in one explanation are reordered in the other. In some cases this is possible; in others it is not. What matters, instead, is that one can generate an explanation that fits the form of a distinctively mathematical explanation that appears to violate our common-sense norms about the acceptable and unacceptable directions of scientific explanation.

If one is committed to the existence of distinctively mathematical explanations of natural phenomena, then one must find a way to reconcile the directionlessness of many applications of mathematics with the directionality of natural explanations. The kinds of relation described in algebra, geometry, and calculus are directionless; with addition or division, a variable on one side of the equation can be moved to the other side. They have no intrinsic left-right directions; rather, these must be imposed from the outside. This is why Lange's examples of putative distinctively mathematical explanation face a directionality challenge. Each of Lange's examples

can be 'reversed' to yield an argument that appeals to the same mathematical premise and that has the same form as Lange's examples but that would not be counted as an acceptable explanation (absent considerable revision in scientific common-sense). Consider, for example:

> **Reversed Strawberries.** Why doesn't Mary have 23 strawberries? Because she divided her strawberries equally among her three kids (EP) and 23 is indivisible by 3 (MP).

Like Strawberries, Reversed Strawberries can be represented as a deductive argument with both an empirical and a mathematical premise:

> 1. Mary evenly distributed her strawberries among her three kids (EP).
> 2. 23 is indivisible by 3 (MP).
> C. Mom doesn't have 23 strawberries.

From a common-sense perspective, at least, Mary's even-numbered pile of strawberries explains but is not explained by her dividing the pile equally among the children.[103] (And surely the number of children mom had is not explained by her distribution of strawberries today, though a mathematical argument of that sort could be constructed as well.) Note further that the implicit directionality in this explanation is plausibly accounted for by ontic assumptions about the kinds of relations that properly carry explanatory force: i.e., that mom's pile is the cause (the source) of the portions each kid gets. In contrast, the portions do not cause the number of strawberries or the number of children. The trefoil knot example faces a similar reversal:

> **Reversed Trefoil Knot**: Why doesn't Terry have a trefoil knot in his shoelace? Because Terry untied the knot (EP) and the trefoil knot is not isotopic to the unknot in three dimensions, and only knots isotopic to the unknot in three dimensions can be untied (MP).

But it would seem more in line with scientific common-sense to explain why Terry has a particular kind of knot by describing how he tied it and not by describing his ability or inability

[103] Catherine Stinson (personal communication) emphasizes that this claim must be bracketed to nonintentional contexts. Mary might decide, for example, to bake a certain number of cookies knowing they will have to be evenly divided among her kids, or she might decide to have three kids because she decides that three is the maximum number of children she can support on her income. These are intentional, causal explanations.

to untie it.

> **Reversed Königsberg**: Why did either zero or two of Königsberg's landmasses have an odd number of bridges in 1756? Because Marta walked through town, hitting each bridge exactly once (EP) and only networks containing zero or two nodes with an odd degree contain an Eulerian path (MP).

As in the other examples, Königsberg's layout is arguably better explained by the decisions of the Burgermeister than by Marta's walk, yet facts about Königsberg's layout follow reliably from descriptions of either.

> **Reversed Chopsticks**: Why were the chopsticks tossed non-randomly? Because more of the tossed chopsticks were oriented vertically than horizontally (EP) and there are more ways for a chopstick to be horizontal than to be vertical (MP).

In this "reversal," the unexpected number of vertically oriented chopsticks provides evidence that some biasing force must be acting upon them (much as deviations from the Hardy-Weinberg equilibrium detect selective forces). As in Lange's forward-directed version of the example, the argument here is inductive. But while we are apt to count Lange's original example as explanatory, it seems more fitting with scientific common-sense to describe Reversed Chopsticks as describing an evidential, not explanatory, relation. In the case of Cicadas, suppose that a field scientist discovers a species of Cicadas that thrives despite the fact that its life cycle overlaps considerably with that of its periodic predators:

> **Reversed Cicadas**: Why doesn't it minimize predation in these Cicadas to have a life cycle that intersects only infrequently with that of your periodic predators? Because cicadas with prime life-cycle periods don't tend to suffer less from predation by predators with periodic life cycles than do cicadas with composite periods (EP) and because prime periods minimize the frequency of intersection (MP).

To modify the example and give a more intuitive appeal, suppose that the life-cycles of a species of Cicada and its periodic predator overlap only every 21 years. This places constraints on the space of possible periods for the life-cycles in these species: 1, 3, 7, and 21 years are the

available options. If we know on empirical grounds that 1 and 21 are not live options and that the life-cycle of the cicada is 7 years, and we package that into the the request for explanation, we can infer with mathematical certainty that the predator cycle is 3 years. But it would seem that the frequency of intersection is explained by the life-cycles, not that the life-cycles are explained by the frequency of their intersection.

> **Reversed Honeycombs**: Why does this species of honeybee divide its combs into regions of unequal area? Because honeybees use less than the amount of wax they would use to divide their combs into hexagons of equal area (EP) and a hexagonal grid uses the least total perimeter in dividing a planar region into regions of equal area (MP).

But it is a stretch from common-sense to think of the bee's hive-construction as explained by the fact that it uses less wax than a hexagonal grid. (If there were such an explanation, it would be a selectionist, and so causal, explanation on Lange's view [498].) The mathematical premise is directionless, but the explanatory force runs in a preferred direction. And finally:

> **Reversed Pendulum**: Why isn't Patty's pendulum a double pendulum? Because Patty's pendulum doesn't have at least four equilibrium configurations (EP) and any double pendulum's configuration space is a torus with at least four stationary points (MP).

But surely Patty's engineering explains the kind of pendulum she has or does not have better than does fact that the pendulum has more or fewer than four equilibrium points (again, outside intentional contexts).

Each of Lange's examples can be used to generate a putative distinctively mathematical explanation, with the same mathematical premise and the same form, that few scientists would accept as a genuine explanation. Given that Lange is not aiming to revise radically our scientific common-sense ideas about the nature of scientific explanation, it would appear that Lange's model of distinctively mathematical explanation is inadequate.

To amplify this point, note that each example of reversal seems to confuse justification

and explanation (see Hempel's [1965] distinction between reason-seeking and explanation-seeking why-questions). An argument justifies believing thesis P (at least partially) when it provides evidence that P. The pristine form of the covering-law model, i.e., one conjoined to the strongest form of the explanation-prediction symmetry thesis, can be seen as attempting to erase this boundary. The goal was to cast explanation as fundamentally an epistemic achievement: explanation is reduced to rational expectation. The problem, of course, is that one can have reason to believe P without explaining P. An Archaeopteryx fossil gives one reason to believe that Archaeopteryx once existed, but it does not explain Archaeopteryx's existence. The same point has been made time and again: with barometers and storms, spots and measles, yellow fingers and lung cancer, and roosters and sunrises. Indicators are not always explainers. It was in recognition of this problem that defenders of the covering-law model quickly backed away from strong forms of the explanation-prediction symmetry thesis and sought other means to account for the directionality of scientific explanations. It was in the face of these challenges that Salmon raised his flag in favor of the ontic conception.

Yet precisely the same problem appears to arise for Lange's examples: We learn something about Terry's knot when we learn he's untied it; we learn something about Königsberg from Marta's stroll; we learn something about our chopsticks when we observe their contra-normal behavior; we learn something about the structure of honeycombs from the amount of wax used; we learn something about the life-cycles of cicadas when we observe predation patterns; and we learn something about a pendulum from how many equilibrium configurations it has. But learning something about the system is not in all cases tantamount to explaining that feature of the system.

Lange argues that the order of explanatory priority in his examples follows the degree of modal necessity, with more necessary things explaining less necessary things. Yet this restriction on distinctively mathematical explanations cannot block the above examples. After all, the same mathematical laws are involved in the forward and reversed cases. We have simply changed the empirical facts. The problem appears to be that the mathematics in these examples is sufficiently flexible about what goes on the right and left hand side of an equation that it doesn't seem to have the resources internal to it to account for the directionality enforced in scientific common-sense. Some extra ingredient is required to sort genuine mathematical explanations from pretenders and, specifically, to sort explanation from justification. In other words, these putative cases of distinctively modal, mathematical explanations of natural phenomena appear to retain an ineliminable ontic component, perhaps working implicitly in the background, but required to account for the preferred direction to the explanation. Mom's pile explains the kids' allotment, and not vice versa, because the allotment is produced from pile. The trefoil knot explains the failure to untie it, and not vice versa, perhaps because structures constrain functions and not vice versa. Similarly, the structure of Königsberg explains which walks are possible around town, but the walks do not explain the structure of the town. Perhaps the movement of the sticks does not explain the forces acting on the sticks because the pattern in the sticks is not causally relevant to the forces acting upon them. Perhaps life-cycle periods explain predation patterns, and not vice versa, because the length of a life-cycle period is causally relevant to the amount of predation. Perhaps the structure of honeycombs explains the amount of wax used, but not vice versa, because the structure of a honeycomb determines the amount of wax needed to build it. And perhaps the shape of Patty's pendulum is explained by her desires in choosing it and not by the

fact that it does or does not have four stable equilibrium points precisely because Patty's desires are causally relevant and (in most non-intentional contexts) the four equilibrium points are not. In other words, in each case, it would appear that various ontic assumptions about what can explain what are called upon to sort out the appropriate direction of the explanation and to weed out inappropriate applications of the same argumentative forms appealing to the same mathematical laws.[104]

The dialectical situation might be put expressed as a tension between three propositions: first, that there are distinctively mathematical explanations of natural phenomena; second; that mathematical explanations are directionless; and third, that explanations of natural phenomena are not directionless.

To resolve this tension, one might deny the first of these propositions, holding that all distinctively mathematical explanations of natural phenomena have at least implicit within them a set of ontic commitments that account for the directionality of the explanations and so for the norms that sort good from bad mathematical explanations. Perhaps once the explanandum has been narrowed to the point that it is susceptible of a distinctively mathematical explanation, the explanandum has been transformed into a mathematical rather than a natural fact. Our above discussion is consistent with this view but in no way forces it upon us. One might also deny that mathematics is directionless. Perhaps some areas of mathematics enforce a direction that corresponds to the explanatory norms in a given domain. This appears not to be the case in Lange's examples, but it does not follow that there are no such cases. Perhaps, that is, there are

[104] Aggregative explanations apply to constitutive relations but exhibit a preferred direction. The mass of the pile of sand is explained by summing the masses of the individual grains. But one can infer the mass of an individual grain from the mass of the whole and the mass of the other grains. This aggregative explanation appears to have the same simple mathematical structure as Strawberries. In this case, it is a constitutive (not causal) relation that apparently accounts for the preferred direction. Perhaps parts explain wholes and not vice versa: an ontic commitment.

distinctively mathematical explanations of natural phenomena that do not face a directionality problem (Huneman, personal communication).

Finally, one might reject the third proposition and allow that explanations of natural phenomena are directionless. This is the extreme caricature of the covering-law model we mentioned above, one that holds to the strong form of the prediction-explanation symmetry thesis. This option involves biting the bullet and accepting that shadows explain flagpoles, that spots explain measles, and that yellow fingers explain lung cancer. (Railton [1981], for example, includes such things in his 'ideal explanatory text'.)

Even if one is tempted to give up on the first proposition and to deny that there truly are distinctively mathematical explanations of natural phenomena, Lange's discussion highlights an important feature of causal and mechanistic explanation that has thus far received very little attention: namely, that all mechanisms are constrained to work within the space of logical and mathematical possibility. If how something works is explained by revealing constraints on its operation (as Craver and Darden [2013], for example, appear to suggest), then one cannot neglect these modal constraints in a complete understanding of mechanistic explanation. In our view, that thesis is interesting enough even if there are not distinctively mathematical explanations of natural phenomena.

7.4. Presuppositions and Constitutive Contexts.

Although we have modeled our reconstructions on Lange's discussion, in which he explicitly states that contingent, empirical facts are part of the explanantia (506), he may object to the form of our examples. He considers and rejects the following pseudo-explanation:

> Why are all planetary orbits elliptical (approximately)? Because each planetary orbit is (approximately) the locus of points for which the sum of the distances from two fixed

> points is a constant [EP], and that locus is (as a matter of mathematical fact) an ellipse [MP]. (508)

Like the previous examples, this one has an empirical premise and a mathematical premise. This is not a distinctively mathematical explanation, according to Lange, because "the first fact to which it appeals [i.e. EP] is neither modally more necessary than ordinary causal laws nor understood in the why question's context to be constitutive of being a planetary orbit (the physical arrangement in question)" (508). However, if we presuppose that the planetary orbits in question are just those that are loci of points for which the sum of the distances from two fixed points is a constant, then that fact is understood in the why question's context to be constitutive of being a planetary orbit. The why-question then becomes: Why are all planetary orbits that are loci of points for which the sum of the distances from two fixed points is a constant, elliptical? It is constitutive of the planetary orbits in question that they are loci of points for which the sum of the distances from two fixed points is a constant. The distinctively mathematical explanation is that those loci are necessarily ellipses. Should Lange object to our "reversed" examples on similar grounds, their empirical premises can also be presupposed and shifted into their associated why-questions. For example, in Reversed Trefoil Knot, instead of asking, "Why doesn't Terry have a trefoil knot in his shoelace?" and stating as an empirical premise that Terry untied the knot, we could instead ask, "Why doesn't Terry have a trefoil knot in the shoelace he untied?" Now the former empirical premise is part of the constitutive context of the why-question. We presuppose that Terry untied his shoelace, rather than stating it as an empirical premise. This seems to fit Lange's criteria for distinctively mathematical explanation.

Lange could respond to this move by distinguishing between what is *understood* to be constitutive of the physical task or arrangement at issue and what is *actually* constitutive of the

physical task or arrangement at issue.[105] Lange could then argue that, for example, in Trefoil Knot it is actually constitutive of the physical task or arrangement at issue that Terry's shoelace is a trefoil knot. However, Lange could continue, in the version of Reversed Trefoil Knot where we presuppose that Terry untied his shoelace, that fact is not actually constitutive of the physical task or arrangement at issue. We are unsure how this distinction between what is "understood" to be and "actually" constitutive could be drawn. When we request an explanation for the fact that Terry failed to untie his shoelace, we grant that context determines that it is actually (and not merely understood to be) constitutive of that fact that his shoelace is a trefoil knot. However, when we request an explanation for the fact that Terry doesn't have a trefoil knot in the shoelace he untied, it seems to us constitutive of that very fact that Terry untied his shoelace. It wouldn't be the same explanandum had Terry not untied his shoelace. We do not see how one can claim that Terry's untying the knot is merely understood to be constitutive of this explanandum, while claiming that the shoelace's being a trefoil knot is actually constitutive of the former explanandum.

We don't think there's anything objectionable about so restricting the range of our explananda/why-question (e.g., to just those planetary orbits that are loci of points for which the sum of the distances from two fixed points is a constant). Notice that such a restriction is required of Lange's examples as well. For example, it is not constitutive of all shoelaces that they contain trefoil knots; it is constitutive only of the shoelace under consideration, which actually contains a trefoil knot. Nor is it constitutive of all pendula that they are double pendula; nor of all arrangements of strawberries and children that there are 23 of the former and 3 of the latter; nor

[105] We thank an anonymous referee for this suggestion and careful discussion of the points in this section. Note that Lange (2013b) always speaks of what is "understood" to be constitutive in the context of the why-question (e.g., 491, 497, 506, 507, 508).

of all bridges that they have a non-Eulerian structure. This response to our challenge, in other words, requires an account of how context determines what is constitutive of the physical task or arrangement in question[106], especially if it relies on a distinction between what is actually and merely understood to be constitutive in a given context.

7.5. Conclusion: Modal and Ontic Aspects of Mechanistic Explanations.

Return again to the flagpole and the shadow. As discussed above, Bromberger and Salmon used this example to demonstrate the directionality of scientific explanations. They enlist this point to argue for an ineliminable causal (or more broadly, ontic) component in our normative analysis of scientific explanation. We have used the same strategy to argue for an ineliminable ontic component in Lange's examples of distinctively mathematical explanation. But the example can be yoked for another duty.

One might, in fact, describe the flagpole example as a distinctively mathematical (or at least trigonometric) explanation of a natural phenomenon, one that calls out for a distinctively modal interpretation. *Presupposing* that the angle of elevation of the sun is θ and that the height of the flagpole is *h* (and the flagpole and ground are straight and form a right angle, and that the system is Euclidean, etc.; EP), why is the length of the flagpole's shadow *l*? Once the contingent causal facts are presupposed in our empirical premise, the only relevant fact left to do the explaining seems to be the trigonometric fact that $\tan \theta = h/l$ (MP). Moreover, once these natural facts are presupposed, the length of the flagpole's shadow seems to follow by trigonometric necessity. So if we package all the natural facts into an empirical premise and highlight the relation $\tan \theta = h/l$, which is crucial for the argument to work, then we might see this as a case in which the bulk of the explanatory force is carried by a trigonometric function. The example thus

[106] This worry is raised by Pincock (2015: 875). We thank an anonymous reviewer for bringing this to our attention.

seems to provide a recipe for turning at least some mechanistic explanations into distinctively mathematical explanations: simply package all of the empirical conditions, such as the rectilinear propagation of light, or the Euclidean nature of spacetime, into the empirical premise or the context of the request for explanation, and leave a mathematical remainder or a tautology to serve as the premise with stronger-than-natural necessity.[107]

The importance of geometry to mechanistic explanation is readily apparent in artifacts, such as the coupling between an engine and the drive crank shaft of a car. Machamer, Darden and Craver (2000) describe the organization of such mechanisms as geometrico-mechanical in nature. Vertical motion produced by explosions in the piston chambers drive the pistons out. The center of each piston is connected via a rod to the crankshaft at some distance (r) from the center of the crankshaft so that when the piston is driven out, the crankshaft is rotated in a circle. This mechanism very efficiently transfers the vertical force of the pistons into a circular motion that drives the car forward. These engine parts are organized geometrically in circles and triangles. The angle of the connecting rod, for example, determines the position of the piston, though the explanation would appear to work the other way around. Yet these mathematical facts surely are relevant to why the car accelerates as it does and not faster or slower.[108]

[107] This could presumably be done with any kind of necessity. For example, take an explanation one of whose premises is a *conceptual* necessity. Fix or presuppose all the premises other than the conceptual necessity. You then have a distinctively *conceptual* explanation. Lange appears to recognize this possibility (504).

[108] Baron, Colyvan, and Ripley (forthcoming; see also Chirimuuta forthcoming) propose assimilating this mathematical dependence to a "counterfactualist" account of explanation (i.e. an account according to which explanatory power consists in the ability to answer what-if-things-had-been-different questions or w-questions) and they show how to assess the relevant counterpossible counterfactuals within a structural equation modeling framework. We find this assimilation plausible but as yet inadequate, because Baron et al. (and Chirimuuta) do not address the question of which true counterfactuals are explanatorily relevant and which are not. For example, there are contexts in which it is true that had the flagpole's shadow been length *l* then the flagpole's height would have been *h*. There are also contexts in which it is true that had Mary divided her strawberries evenly among her children, then 23 would have been divisible by 3. Thus there is a similar problem of directionality with respect to counterfactuals: in one direction, a counterfactual can seem explanatory; in the other direction, it does not seem explanatory. We think that the distinction between explanatorily relevant and irrelevant counterfactuals must be made by appeal to ontic considerations (Salmon 1984; Povich forthcoming a).

Note that if the counterfactualists are right, this will go some way to dissolving the distinction between

But as Lange's examples aptly illustrate, mathematics appears to play an essential role in mechanistic explanations in at least many areas of science. After all, the space of possible mechanisms is constrained by the space of mathematical (and logical) possibility. If one considers the mechanisms of sound transduction in the inner ear, one finds an arrangement most similar to the engine and the crankshaft, except in this case the mechanism converts vibrations in the air into vibrations in fluid. Still, parts are arranged geometrically. Likewise, when we look into the intricate mechanisms gating ion channels, we seem to find structures that are understood geometrically, in terms of sheets and helices, which structures allow or prohibit certain activities (Kandel *et al*. 2013). Structural information has been essential to understanding the mechanisms of protein synthesis and inheritance and to understanding features of macro evolution (Craver and Darden 2013). Perhaps not all of these explanations are distinctively mathematical, but the mathematics does ineliminable work in revealing how the mechanism operates, how it can operate, and how it cannot.

This blend of the mathematical and the mechanical (or more broadly, the ontic) is, in fact, precisely what one would expect based on the history of the mechanical philosophy. Aristotle's mechanics (De Groot 2008) works fundamentally by reducing practical problems to facts about circles. Hero of Alexandria and Archimedes, though celebrated for the practical utility of their

---

ontic and modal conceptions of explanation. According to counterfactualists, causal, mechanistic, distinctively mathematical, and all other kinds of explanation derive their explanatory power from their ability to answer w-questions about their explananda. No one, as far as we know, takes the distinction between causal and mechanistic explanation to be significant enough to warrant relegating each to a different conception of explanation. The distinction between them is real and there is disagreement about how to make it, but, even noting the real differences between causal and constitutive relevance, no one takes the distinction to mark two wholly different conceptions of what it means to explain. If the counterfactualists are right, the distinction between distinctively mathematical explanations and causal/mechanistic explanations seems as insignificant for the theory of explanation as the distinction between causal and mechanistic explanation. There is no philosophically significant reason to lump a few kinds of explanation together and say that they explain in accordance with an "ontic conception" and the others in accordance with a "modal conception". For the counterfactualist, all are simply species of a genus, and all explain by providing answers to w-questions.

simple machines, viewed those machines equally as geometrical puzzles to be solved. Descartes' conception of the mechanistic structure of the world was directly connected with his planar representation of geometrical space, in which extended things interact through contact. Galileo demonstrated his results with thought experiments, such as the Tower of Pisa, that rely on basic mathematical truths (i.e., an object cannot both accelerate and decelerate at the same time). Newton wrote the *Principia*, like the great physicists before him, in the language of geometry. Dijksterhuis (1986) closes his masterly *Mechanization of the Scientific World Picture* with the cautionary note that, "serious misconceptions would be created if mechanization and mathematization were presented as antitheses" (500). It is a misconception because the mathematization of nature and the search for basic mechanistic explanatory principles have been treated historically as distinct aspects of the same explanatory enterprise. The very idea of mechanism, and the idea of the world as a causal nexus, has always been expressed in tandem, rather than in opposition, to the idea that the book of nature is written in the language of mathematics and the belief that a primary aim of science is to leave nothing in words.

Chapter 8. Conclusion and Future Work

8.1 Conclusion

A number of philosophers have agreed that there are models that provide noncausal explanations and that this noncausal explanatoriness has something to do with a model's ability to capture counterfactual dependencies that cannot plausibly be interpreted causally (Batterman 2002a; 2002b; Bokulich 2011; Huneman 2010; Reutlinger 2014; 2016; Rice 2012; 2013; Saatsi and Pexton 2013; Woodward 2013). However, I have argued that this merely provides the germ of a full theory of explanation. To account adequately for explanatory asymmetries, this account requires ontic supplementation. The resulting theory I have called the "generalized ontic conception" (**GOC**), according to which explanatory models provide information about the ontic structures on which the explanandum phenomenon depends.

While an ontic conception, **GOC** recognizes an ineliminable epistemic aspect of explanation: **GOC** makes use of the concept of information, which is relative to background knowledge. However, I have also argued that lay and scientific usage of the concept of explanation suggests that such information should not be relativized to any individual agent. This opens up a novel intersection of philosophy of explanation and social epistemology, though I remain agnostic as to whether there is a social agent that knows the propositions to which information is relativized.

8.2 Future Work

There are at least three areas where *Model and World* needs to be further developed. The first is on the explanation/model distinction. I make this distinction in Chapter 2 and Kaplan and Craver (unpublished) make it in their account of norms of explanatory completeness. They

consider the explanation of some phenomenon to be the total store of knowledge about its causes and mechanisms. This knowledge is not embodied in any single model. Rather, it appears to be contained in something akin to Railton's (1981) hypothetical ideal explanatory text. Individual models capture or express part of this knowledge. If "capture" or "express" means provide information about the contents of the ideal explanatory text (which was Railton's view), then Kaplan and Craver's view seems to be akin to that defended here.

The second is on model semantics. Although **GOC** does not use terms like "reference" or "mapping," this does not imply that model semantics has no place in the theory of explanation. To assess the information conveyed by a model, we see what possibilities are excluded given its *truth*, and model truth will likely involve a concept like reference or similarity.

The third is on the explanatory power of distinctively mathematical explanations and whether or not such explanatory power can be accommodated by **GOC**. When all the contingent, empirical premises of an explanation are presupposed, leaving only a purely mathematical premise, it seems that the only thing on which the explanandum depends is a mathematical fact. Is this a novel kind of ontic dependence between a natural fact and a mathematical fact, perhaps construed along mathematical structuralist lines (Resnik 1997)? Lange's (2017, 42–4) account of explanation by constraint might suggest a kind of ontic dependence – mathematical facts constrain or limit the space of possible causal mechanisms. Or perhaps in such a case the mathematics only reveals negative information – information that the explanandum does not depend on anything other than the presupposed contingent facts? Mathematical trivialism may be helpful for this latter suggestion (Rayo 2010; 2013). According to trivialism, purely mathematical statements make no demands of the world – they require no truthmakers – so their

truth-conditions are trivially satisfied. The modal force of distinctively mathematical explanations may result from their providing information to the effect that the explanandum does not depend on anything.

References

Andersen, H. “Complements, Not Competitors: Causal and Mathematical Explanations.” *The British Journal for the Philosophy of Science* (forthcoming).

Badia, S., F. Guillén-Gónzalez, and J. V. Gutiérrez-Santacreu. “An Overview on Numerical Analyses of Nematic Liquid Crystal Flows.” *Archives of Computational Methods in Engineering* 18 (2011): 285–313.

Baker, A. “Are There Genuine Mathematical Explanations of Physical Phenomena?” *Mind* 114 (2005): 223–38.

Baker, A. and M. Colyvan. “Indexing and Mathematical Explanation.” *Philosophia Mathematica* 19.3 (2011): 323–34.

Baker–Hytch, M. and M. A. Benton. “Defeatism Defeated.” *Philosophical Perspectives* 29.1 (2015): 40–66.

Baron, S., M. Colyvan, and D. Ripley. “How Mathematics Can Make a Difference.” *Philosophers' Imprint* (forthcoming).

Batterman, R. “Multiple Realizability and Universality.” *The British Journal for the Philosophy of Science* 51 (2000): 115–45.

Batterman, R. “Asymptotics and the Role of Minimal Models.” *The British Journal for the Philosophy of Science* 53.1 (2002a): 21–38.

Batterman, R. *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford: Oxford University Press, 2002b.

Batterman, R. “On the Explanatory Role of Mathematics in Empirical Science.” *The British Journal for the Philosophy of Science* 61 (2010): 1–25.

Batterman, R. and C. Rice. “Minimal Model Explanations.” *Philosophy of Science* 81.3 (2014): 349–76.

Baumgartner, M. and A. Gebharter. “Constitutive Relevance, Mutual Manipulability, and Fat-handedness.” *The British Journal for the Philosophy of Science* 67.3 (2016): 731–56.

Bechtel, W. “Looking Down, Around, and Up: Mechanistic Explanation in Psychology.” *Philosophical Psychology* 22.5 (2009): 543–64.

Bechtel, W. and A. Abrahamsen. “Explanation: A Mechanist Alternative.” *Studies in History and Philosophy of the Biological and Biomedical Sciences* 36.2 (2005): 421–41.

Bechtel, W. and R. C. Richardson. *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton: Princeton University Press, 1993.

Beddor, B. "Process Reliabilism's Troubles with Defeat." *Philosophical Quarterly* 65.259 (2015): 145–159.

Beer, R. D. "Toward the Evolution of Dynamical Neural Networks for Minimally Cognitive Behavior." In P. Maes, M. Mataric, J. A. Meyer, J. Pollack, and S. Wilson (eds.), *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*. Cambridge, MA: MIT Press, 1996. 421–9.

Beer, R. D. "The Dynamics of Active Categorical Perception in an Evolved Model Agent." *Adaptive Behavior* 11.4 (2003): 209–43.

Beer, R. D. and P. L. Williams. "Information Processing and Dynamics in Minimally Cognitive Agents." *Cognitive Science* 39.1 (2015): 1–38.

Biederman, I. "Recognizing Depth-rotated Objects: A Review of Recent Research and Theory." *Spatial Vision* 13.2,3 (2000): 241–53.

Biederman, I., E. Cooper, J. Hummel, and J. Fiser. "Geon Theory as an Account of Shape Recognition in Mind, Brain and Machine." In J. Illingworth (ed.), *Proceedings of the 4th British Machine Vision Conference*. London: Springer-Verlag, 1993. 175–86.

Bird, A. "Social Knowing: The Social Sense of 'Scientific Knowledge'." *Philosophical Perspectives* 24.1 (2010): 23–56.

Bird, A. "When is There a Group that Knows? Distributed Cognition, Scientific Knowledge, and the Social Epistemic Subject." In J. Lackey (ed.), *Essays in Collective Epistemology.* Oxford: Oxford University Press, 2014. 42–63.

Bogen, J. "Regularities and Causality; Generalizations and Causal Explanations." *Studies in History and Philosophy of Biology and Biomedical Sciences* 36 (2005): 397–420.

Bokulich, A. "How Scientific Models Can Explain." *Synthese* 180 (2011): 33–45.

Bromberger, S. "Why Questions." In R. G. Colodny (ed.), *Mind and Cosmos*. Pittsburgh: University of Pittsburgh Press, 1966. 86–111.

Brown, J. "Subject-sensitive Invariantism and the Knowledge Norm for Practical Reasoning." *Noûs* 42.2 (2008): 167–89.

Carter, J. A. "Group Knowledge and Epistemic Defeat." *Ergo* 2.28 (2015): 711–35.

Cartwright, N. *How the Laws of Physics Lie*. Oxford: Oxford University Press, 1983.

Chemero, A. *Radical Embodied Cognitive Science.* Cambridge, MA: MIT Press, 2009.

Chemero, A. and M. Silberstein. "After the Philosophy of Mind: Replacing Scholasticism with Science." *Philosophy of Science* 75.1 (2008): 1–27.

Chirimuuta, M. "Minimal Models and Canonical Neural Computations: The Distinctness of Computational Explanation in Neuroscience." *Synthese* 191.2 (2014): 127–53.

Chirimuuta, M. "Explanation in Computational Neuroscience: Causal and Non-Causal." *The British Journal for the Philosophy of Science* (forthcoming).

Cohen, J. and A. Meskin. "An Objective Counterfactual Theory of Information." *Australasian Journal of Philosophy* 84 (2006): 333–52.

Contessa, G. "Scientific Representation, Interpretation, and Surrogative Reasoning." *Philosophy of Science* 74 (2007): 48–68

Couch, M. B. "Mechanisms and Constitutive Relevance." *Synthese* 183.3 (2011): 375–88.

Craver, C. F. "Role Functions, Mechanisms, and Hierarchy." *Philosophy of Science* 68 (2001): 31–55.

Craver, C. F. "Structures of Scientific Theories." In P.K. Machamer and M. Silberstein (eds.), *Blackwell Guide to the Philosophy of Science.* Oxford: Blackwell, 2002. 55–79.

Craver, C. F. "When Mechanistic Models Explain." *Synthese* 153.3 (2006): 355–76.

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

Craver, C. F. "The Ontic Account of Scientific Explanation." In M. Kaiser, O. Scholz, D. Plenge and A. Hüttemann (eds.), *Explanation in the Special Sciences: The Case of Biology and History.* New York: Springer, 2014. 27–54.

Craver, C. F. "The Explanatory Power of Network Models." *Philosophy of Science* (forthcoming).

Craver, C. F. and W. Bechtel. "Top-down Causation without Top-down Causes." *Biology and Philosophy* 22 (2007): 547–63.

Craver, C. F. and L. Darden. *In Search of Mechanisms: Discoveries Across the Life Sciences*.

Chicago: University of Chicago Press, 2013.

Craver, C. F. and J. Tabery. “Mechanisms in Science”. In The Stanford Encyclopedia of Philosophy (Spring 2016 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2016/entries/science-mechanisms/>.

Craver, C. F. and D. M. Kaplan. Unpublished. “Are More Details Better? On the Norms of Completeness for Mechanistic Explanations.

Cummins, R. “Functional Analysis.” *Journal of Philosophy* 72 (1975): 741–65.

Davis, T., B. Love, and A. Preston. “Learning the Exception to the Rule: Model-Based fMRI Reveals Specialized Representations for Surprising Category Members.” *Cerebral Cortex* 22 (2012): 260–73.

Dijksterhuis, E. J. *The Mechanization of the World Picture: Pythagoras to Newton.* Princeton, NJ: Princeton University Press, 1986.

De Groot, J. “*Dunamis* and the Science of Mechanics: Aristotle on Animal Motion.” *Journal of the History of Philosophy* 46.1 (2008): 43–67.

Demir, H. “Counterfactuals vs. Conditional Probabilities: A Critical Analysis of the Counterfactual Theory of Information.” *Australasian Journal of Philosophy* 86 (2008): 45–60.

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

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

Endicott, R. P. “Flat Versus Dimensioned: The What and the How of Functional Realization.” *Journal of Philosophical Research* 36 (2011): 191–208.

Eronen, Markus I. “Levels of Organization: A Deflationary Account.” *Biology and Philosophy* 30.1 (2015): 39–58.

Fisher, R. A. *The Genetical Theory of Natural Selection*. Oxford: Clarendon Press, 1930.

Fodor, J. A. *Psychological Explanation*. New York: Random House, 1968.

Friedman, M. “Explanation and Scientific Understanding.” *The Journal of Philosophy* 1 (1974): 15–19.

Gibson, J. *The Ecological Approach to Visual Perception*. Boston: Houghton-Mifflin, 1979.

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

Giere, R. "How Models Are Used To Represent Reality." *Philosophy of Science* 71 (2004): 742–52.

Glascher, J. and J. O' Doherty. "Model-based Approaches to Neuroimaging: Combining Reinforcement Learning Theory with fMRI Data." *WIREs Cognitive Science* 1 (2010): 501–10.

Glennan, S. "Mechanisms and the Nature of Causation." *Erkenntnis* 44.1 (1996): 49–71.

Glennan, S. "Rethinking Mechanistic Explanation." *Philosophy of Science* 69 (2002): 342–53.

Glennan, S. and P. Illari. (eds.) *The Routledge Handbook of Mechanisms and Mechanical Philosophy* (forthcoming).

Goldenfeld, N. and L. P. Kadanoff. "Simple Lessons from Complexity." *Science* 284.87 (1999): 87–9.

Goldman, A. I. "What is Justified Belief?" In G. S. Pappas (ed.), *Justification and Knowledge*. Dordrecht: Reidel, 1979. 1–25.

Goldman, A. I. "Social Process Reliabilism: Solving Justification Problems in Collective Epistemology." In J. Lackey (ed.), *Essays in Collective Epistemology*. Oxford: Oxford University Press, 2014. 11–41.

Griffith, A. M. "Truthmaking and Grounding." *Inquiry* 57.2 (2015): 196–215.

Grundmann, T. "Reliabilism and the Problem of Defeaters." *Grazer Philosophische Studien* 79.1 (2009): 65–76.

Haken, H., J. A. Scott Kelso, and H. Bunz. "A Theoretical Model of Phase Transitions in Human Hand Movements." *Biological Cybernetics* 51.5 (1985): 347–56.

Halvorson, H. "Scientific Theories." In Paul Humphreys (ed.), *The Oxford Handbook of Philosophy of Science*. Oxford: Oxford University Press, 2016. 585–608.

Hamilton, W. D. "Extraordinary Sex Ratios." *Science* 156.3774 (1967): 477–88.

Harinen, T. "Mutual Manipulability and Causal Inbetweenness." *Synthese* (2014): 1–20.

Harman, G. *Thought*. Princeton: Princeton University Press, 1973.

Hausman, D. *Causal Asymmetries.* New York: Cambridge University Press, 1998.

Hawthorne, J. and J. Stanley. “Knowledge and Action.” *Journal of Philosophy* 105.10 (2008): 571–90.

Hayworth, K. and I. Biederman. “Neural Evidence for Intermediate Representations in Object Recognition.” *Vision Research* 46 (2006): 4024–31.

Heil, J. *From an Ontological Point of View.* Oxford: Oxford University Press, 2003.

Hempel, C. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York, NY: The Free Press, 1965.

Hempel, C. and P. Oppenheim. “Studies in the Logic of Explanation.” *Philosophy of Science* 15 (1948): 135–75.

Hochstein, E. “One Mechanism, Many Models: A Distributed Theory of Mechanistic Explanation.” *Synthese* 193.5 (2016): 1387–407.

Huang, K. *Statistical Mechanics.* New York: Wiley and Sons, 1987.

Huettel, S., A. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging*. Sunderland, Mass.: Sinauer Associates, 2009.

Hummel, J. and I. Biederman. “Dynamic Binding in a Neural Network for Shape Recognition.” *Psychological Review* 99 (1992): 480–517.

Huneman, P. “Topological Explanations and Robustness in Biological Sciences.” *Synthese* 177.2 (2010): 213–45.

Illari, P. “Mechanistic Explanation: Integrating the Ontic and Epistemic.” *Erkenntnis* 78 (2013): 237–55.

Illari, P. and J. Williamson. “What is a Mechanism? Thinking about Mechanisms Across the Sciences.” *European Journal for Philosophy of Science* 2.1 (2012): 119–35.

Izhikevich, E. M. *Dynamical Systems in Neuroscience*. Cambridge, MA: MIT Press, 2007.

Jackson, F. and P. Pettit. “In Defense of Explanatory Ecumenism.” *Economics and Philosophy* 8 (1992): 1–21.

Kaiser, M. I. and B. Krickel. “The Metaphysics of Constitutive Mechanistic Phenomena.” *The British Journal for the Philosophy of Science* (forthcoming).

Kandel, E. R, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth. (eds.) *Principles of Neural Science, 5th Edition*. New York: McGraw-Hill, 2013.

Kaplan, D. M. “Explanation and Description in Computational Neuroscience.” *Synthese* 183 (2011): 339–73.

Kaplan, D. M. “Moving Parts: The Natural Alliance Between Dynamical and Mechanistic Modeling Approaches.” *Biology and Philosophy* 30.6 (2015): 757–86.

Kaplan, D. M. and W. Bechtel. “Dynamical Models: An Alternative or Complement to Mechanistic Explanations?” *Topics in Cognitive Science* 3.2 (2011): 438–44.

Kaplan, D. M. and C. F. Craver. “The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective.” *Philosophy of Science* 78.4 (2011): 601–27.

Kelso, J. A. Scott. “On the Oscillatory Basis of Movement.” *Bulletin of the Psychonomic Society* 18 (1981): 63.

Kelso, J. A. Scott. “Phase Transitions and Critical Behavior in Human Bimanual Coordination.” *American Journal of Physiology: Regulatory, Integrative and Comparative Physiology*, 246.15 (1984): R1000–R1004.

Kim, J. and I. Biederman. “Greater Sensitivity to Nonaccidental than Metric Changes in the Relations Between Simple Shapes in the Lateral Occipital Cortex.” *NeuroImage* 63 (2012): 1818–26.

Kitcher, P. “Explanatory Unification and the Causal Structure of the World.” In W. Salmon and P. Kitcher (eds.), *Minnesota Studies in the Philosophy of Science, Vol 13: Scientific Explanation*. Minneapolis: University of Minnesota Press, 1989. 410–505.

Kitcher, P. and W. Salmon. “Van Fraassen on Explanation.” *The Journal of Philosophy* 84.6 (1987): 315–330.

Krekelberg, B., G. Boynton, and R. J .A. van Wezel. “Adaptation: From Single Cells to BOLD Signals.” *TRENDS in Neurosciences* 29.5 (2006): 250–56.

Kruschke, J. “ALCOVE: An Exemplar-based Connectionist Model of Category Learning.” *Psychological Review* 99 (1992): 22–44.

Lackey, J. “Acting on Knowledge.” *Philosophical Perspectives* 24.1 (2010): 361–82.

Lackey, J. “Socially Extended Knowledge.” *Philosophical Issues* 24.1 (2014): 282–98.

Lackey, J. “What is Justified Group Belief?” *Philosophical Review* 125.3 (2016): 341–96.

Lange, M. “Laws and Meta-laws of Nature: Conservation Laws and Symmetries.” *Studies in History and Philosophy of Modern Physics* 38 (2007): 457–81.

Lange, M. “Conservation Laws in Scientific Explanations: Constraints or Coincidences?” *Philosophy of Science* 78.3 (2011): 333–52.

Lange, M. “Really Statistical Explanations and Genetic Drift.” *Philosophy of Science* 80.2 (2013a): 169–88.

Lange, M. “What Makes a Scientific Explanation Distinctively Mathematical?” *The British Journal for the Philosophy of Science* 64.3 (2013b): 485–511.

Lange, M. “On 'Minimal Model Explanations': A Reply to Batterman and Rice.” *Philosophy of Science* 82.2 (2015): 292–305.

Lange, M. *Because without Cause: Non-causal Explanations in Science and Mathematics*. Oxford: Oxford University Press, 2017.

Lasonen–Aarnio, M. “Unreasonable Knowledge.” *Philosophical Perspectives* 24.1 (2010): 1–21.

Leuridan, B. “Three Problems for the Mutual Manipulability Account of Constitutive Relevance in Mechanisms.” *The British Journal for the Philosophy of Science* 63.2 (2012): 399–427.

Levy, A. “Three Kinds of New Mechanism.” *Biology and Philosophy* 28.1 (2013): 99–114.

Levy, A. “Machine-likeness and Explanation by Decomposition.” *Philosophers' Imprint* 14.6 (2014): 1–15.

Levy, A. and W. Bechtel. “Abstraction and the Organization of Mechanisms.” *Philosophy of Science* 80.2 (2013): 241–61.

Lewis, D. “Causal Explanation.” In David Lewis (ed.), *Philosophical Papers, Vol. 2*. New York: Oxford University Press, 1986. 214–40.

Love, B. C., D. L. Medin, and T. M. Gureckis. “SUSTAIN: A Network Model of Category Learning.” *Psychological Review* 111 (2004): 309–32.

Love, B. C. and T. Gureckis. “Models in Search of a Brain.” *Cognitive, Affective, and Behavioral Neuroscience* 7.2 (2007): 90–108.

Lowe, E. J. “Some Varieties of Metaphysical Dependence.” In M. Hoeltje, B. Schnieder, and A. Steinberg (eds.), *Varieties of Dependence: Ontological Dependence, Grounding, Supervenience, Response-Dependence.* Munich: Philosophia Verlag, 2013. 193–210.

Machamer, P. K., L. Darden, and C. F. Craver. “Thinking About Mechanisms.” *Philosophy of Science* 67.1 (2000): 1–25.

McComb, W. D. *Renormalization Methods: A Guide for Beginners.* Oxford: Clarendon Press, 2004.

Meskin, A. and J. Cohen. “Counterfactuals, Probabilities, and Information: Response to Critics. *Australasian Journal of Philosophy* 86 (2008): 635–42.

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

Morrison, M. “Models as Autonomous Agents.” In Morgan and Morrison (1999). 38–65.

Morgan, M. “Learning from Models.” In Morgan and Morrison (1999). 347–88.

Nagel, E. *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York, NY: Harcourt, Brace, and World, Inc., 1961

Nelson, D. R. “Recent Developments in Phase Transitions and Critical Phenomena.” *Nature* 269.29 (1977): 379–83.

O' Doherty, J., A. Hampton, and H. Kim. “Model-Based fMRI and Its Application to Reward Learning and Decision Making.” *Annals of the New York Academy of Sciences* 1104 (2007): 35–53.

Paicu, M. and A. Zarnescu. “Energy Dissipation and Regularity for a Coupled Navier–Stokes and Q-Tensor System.” *Archive for Rational Mechanics and Analysis* 203 (2012): 45–67.

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

Piccinini, G. and C. F. Craver. “Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches.” *Synthese* 183.3 (2011): 283–311.

Piccinini, G. and C. Maley. “The Metaphysics of Mind and the Multiple Sources of Multiple Realizability.” In M. Sprevak and J. Kallestrup (eds.), *New Waves in the Philosophy of Mind*. Palgrave Macmillan, 2014. 125–152.

Pincock, C. *Mathematics and Scientific Representation.* Oxford: Oxford University Press, 2011.

Pincock, C. “Abstract Explanations in Science.” *The British Journal for the Philosophy of Science* 66.4 (2015): 857–882.

Povich, M. “Mechanisms and Model-based Functional Magnetic Resonance Imaging.” *Philosophy of Science* 82.5 (2015): 1035–46.

Povich, M. “Minimal Models and the Generalized Ontic Conception of Scientific Explanation.” *The British Journal for the Philosophy of Science* (forthcoming a).

Povich, M. “Mechanistic Explanation in Psychology.” In H. Stam and H. Looren de Jong (eds.), *The SAGE Handbook of Theoretical Psychology.* Forthcoming b.

Povich, M. and C. F. Craver. “Mechanistic Levels, Reduction, and Emergence.” In S. Glennan and P. Illari (eds.), *The Routledge Handbook of Mechanisms and Mechanical Philosophy.* Forthcoming.

Priestley, E. B., P. J. Wojtowicz, and P. Sheng. *Introduction to Liquid Crystals.* New York and London: Plenum Press, 1975.

Railton, P. “A Deductive-Nomological Model of Probabilistic Explanation.” *Philosophy of Science* 45 (1978): 206–26.

Railton, P. “Probability, Explanation, and Information.” *Synthese* 48.2 (1981): 233–56.

Rayo, A. “Towards a Trivialist Account of Mathematics”. In O. Bueno and Ø. Linnebo (eds.), *New Waves in Philosophy of Mathematics.* Basingstoke: Palgrave Macmillan, 2010. 239-62.

Rayo, A. *The Construction of Logical Space.* Oxford: Oxford University Press, 2013.

Resnik, M. D. *Mathematics as a Science of Patterns*. Oxford: Clarendon Press, 1997.

Reutlinger, A. “Why Is There Universal Macrobehavior? Renormalization Group Explanation as Noncausal Explanation.” *Philosophy of Science* 81.5 (2014): 1157–70.

Reutlinger, Alexander. “Is There A Monist Theory of Causal and Noncausal Explanations? The Counterfactual Theory of Scientific Explanation.” *Philosophy of Science*, 83 (2016): 733–45.

Rice, C. “Optimality Explanations: A Plea for an Alternative Approach.” *Biology and Philosophy* 27 (2012): 685–703.

Rice, C. “Moving Beyond Causes: Optimality Models and Scientific Explanation.” *Noûs* 49.3 (2015): 589–615.

Romero, F. “Why There Isn’t Inter-level Causation in Mechanisms.” *Synthese* 192.11 (2015): 3731–55.

Ross, Lauren N. “Dynamical Models and Explanation in Neuroscience” *Philosophy of Science*

81.1 (2015): 32–54.

Ruben, D.-H. *Explaining Explanation.* London and New York: Routledge, 1990.

Saatsi, J. and M. Pexton. “Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Noncausal Explanations.” *Philosophy of Science* 80 (2013): 613–24.

Salmon, W. “Statistical Explanation.” In Wesley Salmon (ed.), *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh Press, 1971. 29–87.

Salmon, W. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press, 1984.

Salmon, W. “Four Decades of Scientific Explanation.” In W. Salmon and P. Kitcher (eds.), *Minnesota Studies in the Philosophy of Science, Vol 13: Scientific Explanation*. Minneapolis: University of Minnesota Press, 1989. 3–219.

Sansom, R. and J. Shields, J. “Asymmetry in the Unificationist Theory of Causal Explanation.” *Synthese* (2016): 1–19.

Scarantino, A. “Shell Games, Information, and Counterfactuals.” *Australasian Journal of Philosophy* 86 (2008): 629–34.

Scarantino, A. “Information as a Probabilistic Difference Maker.” *Australasian Journal of Philosophy* 93 (2015): 419–43.

Schmidt, R. C., C. Carello, and M. T. Turvey. “Phase Transitions and Critical Fluctuations in the Visual Coordination of Rhythmic Movements Between People.” *Journal of Experimental Psychology: Human Perception and Performance* 16.2 (1990): 227–47.

Scriven, M. “Explanation and Prediction in Evolutionary Theory.” *Science* 130.3374 (1959): 477–82.

Scriven, M. “Causation as Explanation.” *Noûs* 9.1 (1975): 3–16.

Skow, B. “Are There Non-Causal Explanations (of Particular Events)?” *The British Journal for the Philosophy of Science* 65.3 (2014): 445–67.

Sober, E. “Equilibrium Explanation.” *Philosophical Studies* 43 (1983): 201–10.

Sober, E. *The Nature of Selection*. Cambridge, MA: MIT Press, 1984.

Stepp, N., A. Chemero, and M. T. Turvey. “Philosophy for the Rest of Cognitive Science.” *Topics in Cognitive Science* 3.2 (2011): 425–37.

Strevens, M. *Depth: An Account of Scientific Explanation*. Cambridge: Harvard University Press, 2008.

Strevens, M. “No Understanding without Explanation.” *Studies in History and Philosophy of Science* 44 (2013): 510–15.

Suárez, M. “An Inferential Conception of Scientific Representation.” *Philosophy of Science* 71 Supplement (2004): S767–779.

Suppe, F. *The Structure of Scientific Theories*. Chicago: University of Illinois Press, 1977.

Suppes, P. “A Comparison of the Meaning and Use of Models in the Mathematical and Empirical Sciences.” In H. Freudenthal (ed.), *The Concept and Role of the Model in Mathematics and Natural and Social Sciences*. Dordrecht: Reidel, 1961. 163–77.

Suppes, P. “What is a Scientific Theory?” In S. Morgenbesser (ed.), *Philosophy of Science Today.* New York: Basic Books, 1967. 55–67.

Thalos, M. “Explanation is a Genus: An Essay on the Varieties of Scientific Explanation.” *Synthese* 130 (2002): 317–54.

Tononi, G. and C. Koch. “Consciousness: Here, There and Everywhere?” *Philosophical Transactions of the Royal Society B* **370.1668** (2015).

van Fraassen, B. C. *The Scientific Image*. Oxford: Oxford University Press, 1980.

van Gelder, T. “The Dynamical Hypothesis in Cognitive Science.” *Behavioral and Brain Sciences* 21.5 (1998): 615–28.

van Gelder, T. and R. F. Port. “It's About Time: An Overview of the Dynamical Approach to Cognition.” In R. F. Port and T. van Gelder (eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: The MIT Press, 1995. 1–43.

Walmsley, J. “Explanation in Dynamical Cognitive Science.” *Minds and Machines* 18.3 (2008): 331–48.

Weisberg, M. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press, 2013.

Weiskopf, D. A. “Models and Mechanisms in Psychological Explanation.” *Synthese* 183.3 (2011a): 313–38.

Weiskopf, D. A. “The Functional Unity of Special Science Kinds.” *The British Journal for the Philosophy of Science* 62 (2011b): 233–58.

Weiskopf, D. A. “The Explanatory Autonomy of Cognitive Models.” In David M. Kaplan (ed.), *Integrating Psychology and Neuroscience: Prospects and Problems*. Oxford: Oxford University Press, forthcoming.

West, G., J. H. Brown, and B. Enquist. “The Fourth Dimension of Life: Fractal Geometry and Allometric Scaling of Organisms.” *Science* 284 (1999): 1677–9.

White, C. and R. Poldrack. “Using fMRI to Constrain Theories of Cognition.” *Perspectives on Psychological Science* 8.1 (2013): 79–83.

Williamson, T. *Knowledge and its Limits*. Oxford: Oxford University Press, 2000.

Woodward, J. *Making Things Happen*. Oxford: Oxford University Press, 2003.

Woodward, J. “Mechanistic Explanation: Its Scope and Limits.” *Proceedings of the Aristotelian Society Supplementary Volume* 87 (2013): 39–65.

Wright, C. D. “Mechanistic Explanation Without the Ontic Conception.” *European Journal of Philosophy of Science* 2.3 (2012): 375–94.

Ylikoski, P. and J. Kuorikoski. “Dissecting Explanatory Power.” *Philosophical Studies* 148.2 (2010): 201–19.

Zednik, C. “The Nature of Dynamical Explanation.” *Philosophy of Science* 78.2 (2011): 238–63.

Zednik, C. “Heuristics, Descriptions, and the Scope of Mechanistic Explanation.” In C. Malaterre and P.-A. Braillard (eds.), *Explanation in Biology: An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences.* Dordrecht: Springer, 2015. 295–318.