Causation, Mechanism and Mind

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Daniel Pearlberg

Graduate Program in Philosophy

The Ohio State University

2015

Dissertation Committee:

Richard Samuels, Advisor

Timothy Schroeder

Declan Smithies
Copyright by

Daniel Pearlberg

2015
Abstract

Philosophers of mind and cognitive science have recently gravitated towards a new mechanistic approach to constitutive explanation, and an interventionist approach to causation and causal explanation. In this dissertation I discuss the implications of these new approaches for four issues in the philosophy of mind and cognitive science: Mental causation, the nature of explanation, the extended mind hypothesis, and the dynamicist approach to explanations in cognitive science. I argue that the interventionist account of causation can be used to solve the problem of causal exclusion for non-reductive physicalist approaches to mental causation. However, in order to do so I propose a new improved version of interventionism which rules out the possibility of a certain kind of overdetermination, thereby showing that at most only one of the two different interventionist solutions to the exclusion problem is indeed a viable solution on behalf of non-reductive physicalism—namely, the interventionist proportionality argument. Next, I argue that the ontic and epistemic construals of mechanistic explanations should be seen as reflecting different (though related) legitimate senses of explanation, and there is no reason to think that the different senses of explanation are in conflict with one another. Next, I provide a novel argument against the extended mind hypothesis, employing only premises that proponents of the extended mind hypothesis (including the New Mechanists) have independent reasons to accept. I conclude that while the mutual manipulability account of constitutive relevance—embraced by the New Mechanists—may be useful for demarcating the boundaries of mechanisms studied in many of the sciences, it cannot be used to support the extended mind
hypothesis. Finally, I show that two arguments commonly made in the debate between Dynamicists and the New Mechanists are mistaken, and I conclude that there is— or need be— far less theoretical disagreement between Dynamicists and the New Mechanists than is suggested by the debate.
Acknowledgments

I am extremely grateful to my advisor, Richard Samuels, for many years and countless hours of support in numerous guises. Richard has gone far above and beyond any expectations I had concerning the duties of an advisor, and I feel very fortunate that he agreed to take me on as a student so many years ago. I am also fortunate to have had Declan Smithies and Tim Schroeder on my committee. They too have spent many years supporting and encouraging me, and along with Richard they have played vital roles in my intellectual development. Graduate school can be tough at times, but it would have been an absolute nightmare for me without the kindness and generosity of Richard, Declan and Tim.

I am grateful to my parents- Jay and Debbie Pearlberg- for decades of support. And I could not have done this without the love of my family here in Columbus, human and feline: Brittany Scott, Kaylyn Rhoads, the late Billie Millikan Rhoads, Django Rhoads, Franklin Rhoads, Gary Rhoads, Glenda Rhoads, Peter Rhoads, and Nikolai. Thanks!
Vita

1996 ................................................................. Akiva Day School

2001 ................................................................. B.A. Philosophy, University of Michigan

2003 ................................................................. M.S.c. History and Philosophy of Science,

London School of Economics

2006 .................................................................. J.D. Law, University of Michigan

2006 to present .................................................. Graduate Teaching Associate, Department of

Philosophy, The Ohio State University

Fields of Study

Major Field:  Philosophy

Minor Field:  Neuroscience
# Table of Contents

Abstract ................................................................................................................. ii

Dedication .............................................................................................................. iv

Acknowledgments.................................................................................................. v

Vita .......................................................................................................................... vi

Chapter 1: Introduction ........................................................................................ 1

Chapter 2: Interventionism and the Problem of Causal Exclusion .................... 12

Chapter 3: Making Sense of the Different Senses of Explanation ..................... 55

Chapter 4: Reasons, Causes, and the Extended Mind Hypothesis (with Timothy
Schroeder) .............................................................................................................. 89

Chapter 5: Reframing the Debate Between Dynamicists and the New Mechanists...... 118

References ............................................................................................................. 147
Chapter 1: Introduction

Up until roughly the turn of the 21st century, the background theory of explanation that pervaded contemporary philosophy of mind and cognitive science was the Deductive-Nomological [D-N] model of explanation. As its name implies, the D-N model focuses on providing explanations whereby the occurrence of the phenomenon may be deduced from one or more laws of nature. This view of explanation is seldom held by philosophers of mind and cognitive science anymore. In its place, these philosophers have gravitated towards a new mechanistic approach to constitutive explanation, and an interventionist approach to causation and causal explanation. These new approaches to explanation are particularly attractive to philosophers of mind and cognitive science insofar as the new approaches are meant to capture the practices of the “special” sciences- e.g. biology, neuroscience, psychology, economics, etc.- whereas the D-N model tends to idealize legitimate scientific practice to the practices of fundamental physics.

This shift in approaches to explanation brought with it the hope that philosophers could use these new approaches to make progress on relatively longstanding issues in the philosophy of mind and cognitive science. In this dissertation I discuss the implications of these new approaches for four issues in the philosophy of mind and cognitive science: Mental causation, the nature of explanation, the extended mind hypothesis, and the dynamicist approach to explanations in cognitive science.
In Chapter 2 ("Interventionism and the Problem of Causal Exclusion") I discuss the implications of the interventionist account of causation for a debate in the philosophy of mind concerning mental causation. Mental causation has been a prominent, and contentious, issue in the philosophy of mind ever since Princess Elisabeth pressed Descartes to explain how his theory of substance dualism allowed for mental causation. More recently, physicalist philosophers of mind have worried about how non-reductive versions of physicalism can allow for mental causation. The worry— as expressed in the problem of causal exclusion—is that once we have a physicalist worldview, one that rejects ‘spooky’ entities such as ghosts and Cartesian souls, it seems as though all of the causal work in the world is being done by physical properties, and there is no causal work left to be done by mental properties such as beliefs and desires.

Recently, a number of authors have suggested two different ways of using the interventionist account of causation to solve the problem of causal exclusion, and thereby vindicate the causal efficacy of mental properties. According to the interventionist account of causation, to say that X causes Y is, roughly, to say that changing the value of X via an intervention will result in a change in the value of Y, and the resulting change in the value of Y is due only to the change in the value of X. The basic thought underlying the two different interventionist solutions to the exclusion problem is that discussions of the exclusion problem have traditionally failed to articulate the notion of causation presupposed in the setup of the problem. But once we incorporate the interventionist account of causation into the setup of the exclusion problem, the problem goes away— or so, at least, a number of authors have maintained. According to the interventionist proportionality argument, interventionism can be used to show how higher-level properties can be difference-makers insofar as they are proportionate to their effects in a way that their underlying, lower-level properties are not.
According to the interventionist supervenience argument, interventionism can be used to show how the tight relation between mental and physical properties renders any resulting overdetermination unproblematic.

However, Michael Baumgartner argues that rather than being uniquely positioned to solve the exclusion problem on behalf of non-reductive physicalism, interventionists are actually uniquely disadvantaged when it comes to vindicating the causal efficacy of the mental. This is because the interventionist account of causation is ill-equipped to handle cases where the variables at issue—i.e. mental and physical—bear the type of non-causal dependency relation endorsed by non-reductive physicalists. To make matters worse, Baumgartner’s objection can be extended to show that extant versions of interventionism are incapable of vindicating the causal efficacy of all higher-level properties, let alone mental properties.

In this chapter I discuss the costs associated with different versions of interventionism in light of Baumgartner’s objection, and I provide a new, improved version of interventionism capable of vindicating the causal efficacy of higher-level (including mental) properties. The key move that I make is to suggest that extant versions of interventionism err in trying to use the results of interventions on a single variable in order to determine the causal efficacy of a higher-level property. Instead, I argue, we need to use the results of interventions on multiple variables in order to determine the causal efficacy of a higher-level property.

The only cost of my version of interventionism is that it rules out the possibility of a certain kind of overdetermination—supervenient overdetermination—even though this may have seemed to be a genuine possibility. I then argue that rather than being a cost, ruling out this possibility may instead be seen as a strength of interventionism. Finally, given the impossibility of supervenient overdetermination I show that at most only one of the two different
interventionist solutions to the exclusion problem is indeed a viable solution on behalf of non-reductive physicalism—namely, the interventionist proportionality argument.

In Chapter 3 (“Making Sense of the Different Senses of Explanation”) I discuss a debate amongst New Mechanists concerning the nature of explanation. Some New Mechanists argue that mechanistic explanations are best construed ontically—that is, they should be seen as referring to causal forces ‘out there in the world’—while others argue that mechanistic explanations are best construed epistemically—that is, they should be seen as referring to representations of the causal forces. On the face of it, it isn’t clear why we should view the ontic and epistemic construals as being in conflict with one another rather than being two different, yet equally legitimate, facets of scientific explanation. Moreover, from the perspective of trying to make sense of scientific practice, both the ontic and epistemic senses of explanation at least appear to figure prominently in scientific practice: When we compare the relative merits of different explanations for the same explanandum, for example—a practice ubiquitous across the sciences—we are (typically) comparing the relative merits of different representations of the way things in the world work. And unless one adopts an extreme anti-realist interpretation of scientific practice (which none of the participants in the debate profess to do), the referents of the explanatory representations are supposed to be real states of affairs.

Indeed, in this chapter I show that no good arguments have been offered in support of the claim that the different senses of explanation are in conflict with one another. I think that a charitable interpretation of the dispute reveals two quite different possible disagreements: The first, which I shall refer to as the Uniqueness Dispute, presupposes that there is one and only one thing uniquely worthy of being called ‘explanation’, and the dispute then is over what the correct referent of ‘explanation’ actually is. The second, which I shall refer to as the
Fundamentality Dispute, allows for more than one legitimate sense of explanation, but insists that one of the senses of explanation is more fundamental than the other sense(s). The dispute, then, is over which of the senses of explanation is the fundamental sense of explanation.

I reject the main assumption underlying the Uniqueness Dispute - namely, that there is one and only one thing uniquely worthy of being called ‘explanation’- by providing arguments in favor of ‘explanation’ being a polysemous term- that is, a term with multiple (though related) legitimate senses. First I reject several recent linguistic arguments made by Cory Wright in favor of ‘explanation’ only having one literal term, and then I argue that both the ontic and epistemic senses are required in order to handle three intuitively correct claims about explanations that have been discussed in the debate.

Next I turn to the Fundamentality Dispute, where explicit arguments have heretofore only been made on behalf of the ontic construal of mechanistic explanation. Carl Craver argues that the ontic sense of explanation is fundamental in virtue of certain norms of scientific explanation following form the ontic sense. The first problem with Craver’s fundamentality argument is that even if Craver has shown that the ontic sense of explanation is essential for determining some of the norms of scientific explanation, he has failed to show that the ontic sense is fundamental insofar as he has failed to show that the epistemic sense is not essential for determining any of the norms of scientific explanation. The second problem with Craver’s fundamentality argument- and, for that matter, any fundamentality argument revolving around the norms of scientific explanation- is that, on closer inspection, none of the norms of scientific explanation “follow” from one sense of explanation without equally following from the other sense. Thus, I show that it is in fact quite easy to translate claims that seem to follow from one sense of explanation into claims following from a different sense of explanation.
After rejecting the arguments made in the Uniqueness Dispute and the Fundamentality Dispute, I diagnose the debate as deeply confused insofar as it concerns a pseudo problem— one that arises from a failure to recognize a widely appreciated claim about thoughts: Namely, that in many circumstances thoughts are individuated by their referential contents, which in turn needn’t require cognizers at all. I conclude that the ontic and epistemic construals of mechanistic explanations should be seen as reflecting different (though related) legitimate senses of explanation, and there is no reason to think that the different senses of explanation are in conflict with one another.

In Chapter 4 (“Reasons, Causes, and the Extended Mind Hypothesis”, co-written with Timothy Schroeder) we discuss the implications of the new mechanistic account of explanation for a debate in the philosophy of mind and cognitive science concerning the location of the mind and cognition. The traditional, Cartesian view of the mind being housed in the brain has been challenged recently by proponents of the extended mind hypothesis (EMH). According to EMH, the physical vehicles of at least some of a person’s mental states can be located outside of the person’s body. That is, external objects can (at least partly) constitute at least some mental states, or at least some cognitive processes. Thus, according to Andy Clark and David Chalmers’ canonical statement of the view, “beliefs can be constituted partly by features of the environment, when those features play the right sort of role in driving cognitive processes.”

Clark & Chalmers use the following thought experiment to argue for EMH: Inga has a belief concerning the location of New York’s Museum of Modern Art (MoMA) that she can consult by searching her memory. Otto suffers from Alzheimer’s disease and so does not have

---

the luxury of consulting his memory in the ordinary way. Instead, he consults a notebook he carries around with him: a notebook that allows him to replicate certain functional features of ordinary human memory. Clark & Chalmers infamously argue that it is possible that Otto’s case is one in which an inscription in Otto’s notebook, such as the one noting the location of MoMA, has sufficient functional similarities to ordinary memories (such as Inga’s) about the location of MoMA, and so it is possible for Otto to be a person who satisfies EMH.

The debate over EMH appears to have reached a stalemate, insofar as the criticisms rely upon premises not shared by proponents of EMH. One recent suggestion, then, by David Michael Kaplan, has been to use the New Mechanists’ mutual manipulability (MM) account of constitutive relevance in order to allow for a non-question begging, and empirically based, means of resolving the debate in favor of EMH. Since the MM account of constitutive relevance can be used to tell us what the relevant components of a given mechanism are, it can be used to solve the boundary demarcation problem for cognition and, at the same time, allow for the possibility of extended cognition: Something should count as part of a cognitive process, or mechanism, if it bears a relation of mutual manipulability to the mechanism as a whole. And if the part satisfies the conditions of the MM account yet happens to be located external to the agent’s body, then EMH will have been empirically confirmed.

After pointing out several challenges facing Kaplan’s proposal, we grant for the sake of argument that Kaplan can find a way to revise his proposal so as to address these challenges. We then provide a novel argument against EMH, employing only premises that proponents of EMH (including New Mechanists) have independent reasons to accept. Thus, our argument against EMH does not depend on any of the particular details of the MM account of constitutive relevance. Nor does it rely on functional disanalogies between Otto and ordinary human beings.
that are just a matter of fine detail or degree. Instead, our argument makes use of the following necessary condition on holding a belief for a reason: To believe something for a reason one must come to believe it in virtue of attitudes that rationalize— that is, stand in epistemically justifying logical or probabilistic relations to— the belief that is formed, and one must come to believe in virtue of the fact that these attitudes rationalize the new belief.

After clarifying and justifying the necessary condition on holding a belief for a reason, we use it to show that many of the external objects that defenders of EMH have claimed are (in conjunction with an internal cognitive system) beliefs cannot (even partly) constitute beliefs had for a reason. Thus, in the case of Otto, it is not the case that the notebook inscription is there because the attitudes that rationalize the inscription caused it in virtue of rationalizing it. For there is, necessarily, an intermediate step between the attitudes that rationalize the inscription, and the existence of the inscription: Namely, an intentional action consisting of his writing down the location because he wants to remember it. The intentional action severs the direct connection between the rationalizing reason and the notebook inscriptions. In general, then, external states that are generated via intentional action, and not directly in virtue of other attitudes that rationalize them, are necessarily disqualified from being counted as beliefs held for reasons, given the necessary condition on holding a belief for a reason.

Our first thesis, then, is that beliefs had for reasons cannot be extended in the most interesting way specified by EMH. Furthermore, if we add to our argument the premise that every token belief is something that could have been had for a reason, we can derive the stronger thesis that no belief can be extended in this way. We underscore the plausibility of this additional premise by noting how difficult it is to think of cases that, given how they are causally situated, (a) could not have been had for a reason, yet (b) still qualify as beliefs. After
responding to a possible objection, we conclude that while the MM account of constitutive relevance may be useful for demarcating the boundaries of mechanisms studied in many of the sciences, it cannot be used to support EMH.

In Chapter 5 (“Reframing the Debate Between Dynamicists and the New Mechanists”) I discuss the implications of the new mechanistic account of explanation for the dynamicist approach to explanations in cognitive science. According to the New Mechanists, good explanations in cognitive science consist in showing how a cognitive phenomenon is produced by a cognitive mechanism, in virtue of the activities and organization of the underlying parts of the mechanism. For example, a good explanation of spatial memory will involve a description of the spatial memory mechanism that is responsible for the storage of spatial memories, in virtue of the activities, and organization, of the mechanism’s components—e.g. hippocampal neurons, synapses, and NMDA receptors. In contrast, Dynamicists argue that it is dynamical equations, rather than mechanisms, that are doing the explanatory work in (at least some) cognitive science explanations. For example, Dynamicists use the Haken-Kelso-Bunz model’s differential equation to provide an explanation for the changes in the phase relationship between two moving index fingers, and the equation makes no reference to the activities and organization of the underlying parts of the mechanism (if there is one) responsible for bimanual coordination.

In this chapter I first show that two arguments commonly made in the debate between Dynamicists and the New Mechanists are mistaken. The first argument is made by the New Mechanists, against Dynamicists: New Mechanists argue that to the extent that equations in dynamical systems explanations do not refer to the underlying causal structure of cognitive systems, they are not genuinely explanatory, for much the same reason that the D-N model is no longer considered an adequate account of scientific explanation: There is more to (good)
explanation than mere descriptive adequacy and predictive capability. Thus, dynamical systems explanations of cognitive systems may be descriptively accurate, and they may generate a wide array of predictions; but they do not explain how the component parts and operations generate the behavior described by the equations. I show that this argument is mistaken, for we can use the interventionist account of causal explanation to show why Dynamicists needn’t claim that the explanatory norms of predictivism are sufficient for good explanation. Thus, Dynamicist equations can be interpreted as citing genuine difference-making causes featured in relevantly invariant generalizations, and needn’t be interpreted as being merely descriptively, or predictively, adequate.

The second argument is made by Dynamicists against the New Mechanists: Dynamicists argue that dynamical systems explanations are explanatory (at least in part) in virtue of abstracting away from the details of the underlying causal mechanisms, and the explanatory role of abstraction cannot be captured by the New Mechanists. I show that this argument is mistaken, for I use the mutual manipulability account of constitutive relevance to show why even the version of New Mechanism that is the least hospitable to Dynamicists can, and should, embrace the explanatory role of abstraction.

After showing that these two arguments are mistaken, I argue that to the extent that the different approaches are incompatible, the dispute really turns on an empirical issue concerning the causal contributions of the parts of cognitive mechanisms. If it turns out that Dynamicists are right about the empirical issue, then the New Mechanist approach to explanations in cognitive science must be wrong. If, on the other hand, it turns out that the New Mechanists are right about the empirical issue, I show that it doesn’t follow that the Dynamicist approach to explanations in cognitive science must be wrong: First, Dynamicists can be seen as
interventionists addressing a different type of *explanandum*-that is, addressing the question of what causes Y- than the type of *explanandum* addressed by the New Mechanists, who address the question of how does X cause Y. Second, even if the New Mechanists argue that explanations in cognitive science do, or should, aspire to tell us more than just what causes Y, Dynamicists can point out that Dynamicist explanations of tightly coupled systems can pass the explanatorily relevant portions of the MM account of constitutive relevance, and therefore *can*-even by the New Mechanists’ lights- tell us more than just what causes Y. I conclude that there is- or need be- far less theoretical disagreement between Dynamicists and the New Mechanists than is suggested by the debate.
Chapter 2: Interventionism and the Problem of Causal Exclusion

One of the problems that has been raised with regards to the causal efficacy of the mental is the problem of causal exclusion. The problem is typically presented as a reductio of non-reductive physicalism: Jaegwon Kim (1998; 2005) argues that unless mental properties are identical to physical properties, we are forced to conclude that mental events qua mental cannot cause physical events. Since we know that mental events qua mental do cause physical events, it must be the case, pace non-reductive physicalism, that mental properties are identical to physical properties.

Non-reductive physicalists have responded to this challenge in a number of different ways. One recent version of a non-reductive physicalist response has been to focus on the inadequacy of the account of causation presupposed by the exclusion problem. Thus, a number of authors have suggested incorporating the interventionist account of causation into the setup, and thereby dissolving the problem. However, this strategy has not gone unchallenged. Michael Baumgartner (2009; 2010; 2013) argues that rather than being uniquely positioned to solve the exclusion problem on behalf of non-reductive physicalism, interventionists are actually uniquely disadvantaged when it comes to vindicating the causal efficacy of the mental. To make matters

---

2 See, inter alia, Yablo (1992); Shoemaker (2001); Pereboom (2002); Bennett (2003; 2008); Macdonald & Macdonald (2006); Haug (2009; 2010); Wilson (2011); Carey (2010); and Corry (2013).

3 See Raatikainen (2010); Menzies (2008); List & Menzies (2009); Shapiro & Sober (2007); Woodward (2008a; 2014); Shapiro (2010; 2011).
worse, Baumgartner’s objection can be extended to show that extant versions of interventionism are incapable of vindicating the causal efficacy of all higher-level properties, let alone mental properties.

In this chapter I discuss the costs associated with different versions of interventionism in light of Baumgartner’s objection, and I provide a new, improved version of interventionism capable of vindicating the causal efficacy of higher-level (including mental) properties. The only cost of my version of interventionism is that it rules out the possibility of a certain kind of overdetermination, even though this may have seemed to be a genuine possibility. I then argue that rather than being a cost, ruling out this possibility may instead be seen as a strength of interventionism, and it can shed new light on certain aspects of the exclusion problem.

In 2.1 I present the problem of causal exclusion, and in 2.2 I explain why certain types of solutions to the problem should be seen as non-starters. In 2.3 I present an overview of the interventionist account of causation, as well as two different ways in which interventionism has been recruited to solve the exclusion problem. In 2.4 I present Baumgartner’s objection to the interventionist solutions, and I show how the objection can be expanded to show that interventionism is incapable of vindicating the causal efficacy of all higher-level properties. In 2.5 I present James Woodward’s revision of interventionism in light of Baumgartner’s objection, and I argue that Woodward’s revision fails to provide necessary and sufficient conditions for higher-level causation. In 2.6 I provide a new, improved version of interventionism capable of vindicating the causal efficacy of higher-level properties, in 2.7 I argue that the costs associated with my version may instead be seen as a strength, and finally, in 2.8 I use my version of interventionism to shed light on certain aspects of the exclusion problem.
2.1. The Problem of Causal Exclusion

The problem of causal exclusion arises from the inconsistency of the following five theses, each of which is typically accepted by non-reductive physicalists.\textsuperscript{4}

\textit{Distinctness}: No mental properties are identical to physical properties.

\textit{Completeness}: All physical occurrences that have sufficient causes have sufficient physical causes.

\textit{Efficacy}: Some mental properties are sufficient causes of physical occurrences.

\textit{Overdetermination\textsubscript{1}}: Not all physical effects of mental causes are overdetermined by having, in addition to a sufficient mental cause, a sufficient physical cause as well.

\textit{Overdetermination\textsubscript{2}}: If an effect has more than one sufficient cause (at a particular time \(t\)), then the effect is overdetermined.\textsuperscript{5}

Here is how the inconsistency arises:

(P1) Mental property \(M\) is a sufficient cause of physical occurrence \(E\). [\textit{Efficacy}]

(P2) Physical property \(P\) is a sufficient cause of physical occurrence \(E\). [\textit{Completeness}]

(P3) \(M\) is not identical to \(P\). [\textit{Distinctness}]

(P4) \(E\) has more than one sufficient cause. \([(P1), (P2), (P3)]\)

(P5) \(E\) is overdetermined. \([(P4), \textit{Overdetermination\textsubscript{2}}]\)

\textsuperscript{4} This way of presenting the problem is an adaptation of Raatikainen (2010) and Bennett (2007).

\textsuperscript{5} Typically \textit{Overdetermination\textsubscript{2}} is referred to as \textit{Exclusion}, while \textit{Overdetermination\textsubscript{1}} is simply referred to as \textit{Overdetermination}. It will become clear in 2.3.2.2 why I have chosen to use these non-standard labels, though nothing of substance hangs on this choice.
This is a general schema that, given widely held non-reductive physicalist assumptions, will apply anytime (P1) is true—i.e., for any cases in which (P1) is true, (P5) is true as well. This in turn means that (P5) is inconsistent with Overdetermination1—i.e., for any time M causes E, a non-reductive physicalist will be committed to the view that E is overdetermined. The redundancy of mental causes revealed by the exclusion argument only follows, though, given the acceptance of all five theses—Distinctness, Completeness, Efficacy, Overdetermination1, Overdetermination2—and thus Kim uses the exclusion argument as a reductio of non-reductive physicalism.6 A straightforward way to avoid the problem of causal exclusion is to reject Distinctness, which of course the non-reductive physicalist cannot do. Since, says Kim, the refusal to reject Distinctness inevitably leads to the conclusion that mental causation cannot be vindicated, we therefore have a strong reason to reject Distinctness and embrace type-identity.

2.2. Non-Starter Solutions to the Problem of Causal Exclusion

Interventionist solutions to the problem of causal exclusion have focused on Completeness and the Overdetermination premises. In this section I explain why rejecting Distinctness or Efficacy are typically seen as non-starter solutions to the problem.

2.2.1. Supervenience

According to Distinctness, no mental properties are identical to physical properties. Discussion of the exclusion problem has taken place within the context of widespread endorsement of physicalism, and so the contrast to the identity view (mental properties are

---

6 Papineau (2002: 17-18) uses the exclusion argument against dualism in a similar fashion.
identical to physical properties) is (typically) the supervenience view (mental properties supervene upon, but are not identical to, physical properties). Many different versions of supervenience have been discussed in the literature.\(^7\) For present purposes we can think of ‘supervenience’ as indicating a sort of halfway house of dependency - not as strong as the identity dependency relationship, yet not as weak as the dependency relationships endorsed by emergentists and dualists. Thus, by saying that M supervenes upon P [\textit{Supervenience}], I mean to capture the following three claims concerning the relationship between M and P\(^8\):

(a) Every change in the value of M is necessarily accompanied by a change in the value of P, where the relevant modality is metaphysical necessity, and not (just) physical/nomological necessity.

(b) M is not identical to P.

(c) There is no causal relationship between M and P: M is not a cause of P, and P is not a cause of M.\(^9\)

Even though the identity view (reductive physicalism) is typically contrasted with \textit{Supervenience} (non-reductive physicalism), the rejection of \textit{Distinctness} does not entail the truth of \textit{Supervenience}. This is because a third option exists - the emergentist option. I specify the view as follows: M is an emergent property iff

\(^7\) See, e.g. McLaughlin (1995); Bennett & McLaughlin (2005).

\(^8\) Throughout the rest of the paper, when I say that one property supervenes upon another property, unless otherwise indicated I am referring to the supervenience relationship characterized by (a), (b), and (c).

\(^9\) This is standard fare in characterizations of supervenience - perhaps the only prominent dissenter is Searle (1992).
(i) There is some type of dependence relationship between M and P\textsuperscript{10}, but the dependence relationship is weaker than metaphysical necessity; and

(ii) M possesses a novel, fundamental causal power, over and above any of P’s causal powers.

To further clarify the emergentist option, here are some views about the causal powers of M that do not fall under the purview of emergentism (as I am defining it):

- M is a property (or power) had by a whole that is not had by the individual parts of the whole\textsuperscript{11}

- The causal efficacy of M is not metaphysically distinct from the causal efficacy of P, but there is some sort of epistemic disconnect (e.g. in -principle failure of deducibility or predictability) between the causal efficacy of M and the causal efficacy of P.\textsuperscript{12}

- The causal efficacy of M is metaphysically distinct from the causal efficacy of P, but not in virtue of possessing a novel, fundamental causal power. Rather, it is metaphysically distinct in virtue of, say, possessing a proper subset of the powers had by P\textsuperscript{13}, or perhaps in virtue of non-causally determining some of P’s causal powers.\textsuperscript{14}

The reason why interventionism should not reject Distinctness, and should endorse Supervenience with respect to mental properties is because the alternatives are not very plausible views about the relationship between mental and physical properties. Indeed, more

\textsuperscript{10} Hence the idea that M emerges from P.

\textsuperscript{11} This is what Bedau (2008) refers to as ‘nominal’ emergence, and Gillett (2010) refers to as ‘qualitative’ emergence.

\textsuperscript{12} See Wilson (forthcoming). This is also sometimes referred to as epistemological emergentism, in contrast to metaphysical emergentism.

\textsuperscript{13} Wilson (forthcoming) refers to this as ‘weak emergence’ and equates it with the position staked out by NRP.

\textsuperscript{14} See Gillett (2006; 2010). This is probably the most sophisticated version of emergentism around, but it is not what I have in mind, since it claims that the relationship between M and P is one of metaphysical necessity, and thus appears to fall within the purview of NRP.
generally, the alternatives to *Supervenience* with respect to higher-level properties studied across the sciences are not very plausible views about interlevel relations in the sciences. Type-identity is implausible in that there are very few cases in the sciences where we find an identity of interlevel variables. Thus, a quick glance at the special sciences reveals the implausibility of identifying the higher-level properties studied by, say, psychology, with the lower-level properties studied by, say, neuroscience. Take a given psychological property—belief, pain, desire, fear, etc. For each of these properties, there are *many* different possible configurations of neurons that can realize said psychological property. Nor is the implausibility restricted to psychological properties. Indeed, the properties studied in neuroscience *themselves* seem to resist identification with lower-level properties in the same way. For example, for any given ion channel, there are many different configurations of protein subunits that can realize said ion channel.

This problem is, of course, the problem of multiple realization, and starting with Hilary Putnam’s (1967) it has been used as an influential criticism of the type-identity theory in the philosophy of mind, and more generally (starting with Jerry Fodor’s 1974) it has been used to argue for the conclusion that the properties studied by higher-level sciences do not (typically) bear an identity relation with their lower level realizers.15

---

15 This isn’t to say that the multiple realization of higher-level properties has gone unquestioned. Indeed, in recent years the hegemony of the multiple realization argument has been somewhat reduced, in part due to greater attention being paid to the metaphysics of the realization relation, as well as the role played by realization in the sciences. For more recent arguments *against* widespread multiple realization, see Shapiro & Polger (2012). And for more recent arguments on behalf of widespread multiple realization and in response to the critiques, see Aizawa & Gillett (2011).
Emergentism—as defined here— is also implausible, insofar as it has had a poor track record in the sciences for quite some time now.\textsuperscript{16} According to Brain McLaughlin’s (1992) canonical account, emergentism was a scientifically respectable notion up until the quantum mechanical revolution. Before then, there seemed to be big gaps between biology and chemistry, and likewise between chemistry and physics. The gaps were epistemic, but they seemed to provide good evidence in favor of the existence of \textit{metaphysical} gaps. For example, it was posited that the explanation for chemical bonding required the existence of special chemical forces that “emerged” at the level of chemical elements.

All of this, however, changed abruptly with the advent of quantum mechanics, as the perceived gaps between the sciences began to shrink. Thus, McLaughlin writes, Quantum mechanical explanations of chemical bonding suffice to refute central aspects of...Chemical Emergentism. Chemical bonding can be explained by properties of electrons, and there are no fundamental chemical forces. (1992: 89)

Thus, given that epistemic gaps between biology and chemistry, and between chemistry and physics, did not end up warranting the inference that there are corresponding metaphysical gaps, we should be wary of positing metaphysical gaps between the mental and physical on the basis of perceived epistemic gaps between the two. The basic idea, then, is that our best current scientific practices and theories weigh in favor of a relationship between the mental and the physical— and indeed between all higher-level properties studied in the sciences and their lower-level realizers— that is closer than the relationship posited by emergentism, but not as close as the relationship posited by identity theorists. That is, our best current scientific practices and theories weigh in favor of \textit{Supervenience}.

\textsuperscript{16} Of course, since the term ‘emergentism’ has been applied to a number of different positions, whether or not the following objection is valid depends a great deal upon how one is using the term.
Before moving on, I should caution that the overall argument of this chapter does not rest too heavily on the above arguments in favor of *Supervenience*. To begin with, the interventionist solutions to the exclusion problem are typically presented *on behalf of* non-reductive physicalism, and so already *presuppose* the truth of *Supervenience*. More importantly, I will later argue that interventionism should at least countenance the *possibility* of *Supervenience*, even if interventionists think that *Supervenience* is in fact false (notwithstanding the above arguments in favor of *Supervenience*). For now, though, I will be proceeding on the assumption that there are good reasons to think that *Supervenience* is true, and thus the rejection of *Distinctness* is a non-starter.

2.2.2. Efficacy

The rejection of *Efficacy* is typically seen as a non-starter when it comes to solving the exclusion problem, for the rejection of *Efficacy* entails the truth of Type-Epiphenomenalism [Type-E]: Mental events do not cause physical events in virtue of their mental properties. Jerry Fodor famously sounds the alarm bells at the very thought of accepting epiphenomenalism:

“...if it isn’t literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for saying..., if none of that is literally true, then practically everything I believe about anything is false and it’s the end of the world.” (1987: 156)

17 Following the epiphenomenalism literature I will use the terms ‘type’, ‘property’, and ‘kind’ more or less interchangeably. As McLaughlin notes (1989: 29) “Nothing turns on the shift from talk of types [or kinds] to talk of properties. The properties in question are those of falling under event-types. For example, if M is a type of mental event, then the relevant mental property would be the property of being an M.”
More precisely, we can distinguish three main arguments that have been used against Type-E: the common sense objection, the explanation objection, and the ontological objection.$^{18}$

According to the common sense objection, the falsity of Type-E can be demonstrated simply by noting that endorsing Type-E conflicts with our common sense beliefs concerning the causal efficacy of the mental. Perhaps the most explicit version of this objection can be found in Horgan (1989: 47)$^{19}$:

Even if individual mental events and states are causally efficacious, are they efficacious qua mental? I.e., do the mental types (properties) tokened by mental events and states have the kind of relevance to individual causal transactions which allows these properties to figure in genuine causal transactions?...our common-sense belief in the efficacy of the mental presupposes a positive answer to the question just posed; thus a negative answer would constitute a version of epiphenomenalism hardly less offensive to common sense than is the version which denies that mental events have any effects at all.

The explanation objection involves an inference to the best explanation concerning results from psychology$^{20}$- namely, that the best explanation of the success of psychological explanations that attribute causal efficacy to mental phenomena is that the mental is indeed causally efficacious qua mental. Thus, according to Kim,

---

$^{18}$ This taxonomy is taken from Horowitz (1999), though what I am calling here the common sense objection, Horowitz refers to as the introspection objection.

$^{19}$ A similar version of this objection can be found in Kim (1998: 31).

$^{20}$ The objection can be run with results from either scientific psychology or folk psychology- nothing here hinges on this distinction.
It seems plain that the possibility of psychology as a theoretical science capable of generating law-based explanations of human behavior depends on the reality of mental causation: mental phenomena must be capable of functioning as indispensable links in causal chains leading to physical behavior.” (1998: 31)

Finally, according to the ontological objection, Type-E implies mental irrealism- the position that mental properties do not really exist. For if Type-E is true, then mental properties lack causal powers. And, according to Alexander’s Dictum, “to be real is to have causal powers.” (Kim 1993: 202) Thus, if Alexander’s Dictum is true, then if mental properties lack causal powers they don’t really exist, and so mental realism requires the falsity of Type-E.

Interventionists have followed suit in resisting the rejection of Efficacy, typically for reasons similar to the common sense objection and the explanation objection. Thus, given that the rejection of either Distinctness or Efficacy is a non-starter way to solve the exclusion problem, in the next section I shall focus on the remaining theses in the set-up of the exclusion problem- Completeness and Overdetermination- as seen through the lens of the interventionist account of causation.

2.3. The Interventionist Solutions to the Exclusion Problem

A number of authors have argued that the interventionist account of causation can be recruited to solve the problem of causal exclusion. In this section I present two interventionist

21 Another discussion of this type of argument can be found in Fodor (1987).

22 It is named ‘Alexander’s Dictum’ after Samuel Alexander, a leading emergentist of the early twentieth century. Kim importantly adds the following to the above slogan: “I believe this principle, as applied to concrete existents and their properties, will be accepted by most non-reductive physicalists.” (emphasis added)
solutions, each of which is best seen as challenging a particular premise in the setup of the exclusion problem: The proportionality argument, which is best understood as challenging *Completeness*, and the supervenience argument, which is best understood as challenging the overdetermination premises (*Overdetermination1/Overdetermination2*)\(^\text{23}\). First, though, I must digress with a brief overview of the interventionist account of causation.

2.3.1. Brief Overview of the Interventionist Account of Causation

The interventionist account of causation descends from earlier manipulability theories of causation.\(^\text{24}\) The intuitive idea motivating manipulability theories of causation is that if X causes Y, then we can change Y by manipulating X. Moreover, manipulability theories provide a difference-making notion of causation, such that (roughly) a cause must make a difference to its effect in that the presence of the cause results in the presence of the effect, while the absence of the cause results in the absence of the effect.\(^\text{25}\) Early versions of manipulability theories focused on the human agency involved in manipulating causes and effects, and hence were known as agency theories. Dissatisfaction with the anthropocentric account of causation provided by agency theories led to a very different type of manipulability theory - one that divorced causation from human agency, and took as its inspiration recent advances in the causal modeling literature.\(^\text{26}\) I will refer to the more recent type of manipulability theory as the

\(^{23}\) This taxonomy is taken from Raatikainen (2013).

\(^{24}\) E.g. von Wright (1971) and Menzies & Price (1993).

\(^{25}\) The difference-making notion of causation is a *genus* that includes as its *species* the probabilistic, interventionist, and contrastive accounts of causation, as well as specific versions of the counterfactual account of causation. See, *inter alia*, List & Menzies (2009: 5).

\(^{26}\) See Spirtes, Glymour, & Scheines (1993); Pearl (2000).

On Woodward’s version of interventionism, the causal *relata* are represented by variables capable of taking more than one value. Moreover, causal claims made within this framework are made with respect to a particular variable set $V$ under consideration. On this account, to say that $X$ causes $Y$ with respect to $V$ is, roughly, to say that changing the value of $X$ via an intervention, or manipulation, will result in a change in the value of $Y$, and the resulting change in the value of $Y$ is due *only* to the change in the value of $X$ (along, perhaps, with the corresponding change in the value of any other variables in $V$ that lie on the causal path from $X$ to $Y$). Of foremost concern for the interventionist is ensuring that the resulting change in the value of $Y$ is due *only* to the change in the value of $X$. Otherwise, we are not entitled to claim that $X$ causes- and is not merely *correlated* with- $Y$.

More formally, Woodward’s account interdefines the notions of causation and intervention. Here is his account of causation:

---

27 “…variables are properties or magnitudes that, as the name implies, are capable of taking more than one value...Many of the familiar examples of so-called property causation discussed in the philosophical literature may be understood as relationships between two-valued or binary variables, with the variables in question taking one of two values, depending on whether the properties in question are instantiated or not.” (Woodward 2003: 39)

28 A quick aside concerning notation: In the exclusion problem literature, ‘M’, ‘P’, and ‘E’ are typically used to denote, respectively, a mental cause, a physical cause, and the effect; in the interventionism literature ‘X’, ‘Z’, and ‘Y’ are typically used to denote, respectively, a higher-level cause, a lower-level cause, and the effect. This chapter includes direct quotes from both the exclusion problem literature and the interventionism literature, and so I have found it easiest to include both sets of notations rather than trying to rewrite, say, all occurrences of ‘M’ as ‘X’ or vice-versa. Thus, the changes in notation throughout the chapter have no significance other than simply reflecting different notations prevalent in different literatures.
(M) A necessary and sufficient condition for \( X \) to be a (type-level) *direct cause* of \( Y \) with respect to a variable set \( V \) is that there be a possible\(^{29}\) intervention on \( X \) that will change \( Y \) or the probability distribution of \( Y \) when one holds fixed at some value all other variables \( Z_i \) in \( V \). A necessary and sufficient condition for \( X \) to be a (type-level) *contributing cause* of \( Y \) with respect to variable set \( V \) is that (i) there be a directed path from \( X \) to \( Y \) such that each link in this path is a direct causal relationship...and that (ii) there be some intervention on \( X \) that will change \( Y \) when all other variables in \( V \) that are not on this path are fixed at some value. (2003: 59)

And here is his analysis of an intervention:

(IV) I is an intervention variable for \( X \) with respect to \( Y \) iff I meets the following conditions:

(I1) I causes X

(I2) I acts as a switch for all other variables that cause X. That is, certain values of I are such that when I attains those values, \( X \) ceases to depend on the values of other variables that cause \( X \) and instead depends only on the value taken by I.

(I3) Any directed path from I to \( Y \) goes through \( X \). That is, I does not directly cause \( Y \) and is not a cause of any causes of \( Y \) that are distinct from \( X \) except, of course, for those causes of \( Y \), if any, that are built into the I-X-Y connection itself; that is, except for (a) any causes of \( Y \) that are effects of \( X \) (i.e. variables that are causally between \( X \) and \( Y \)) and (b) any causes of \( Y \) that are between I and \( X \) and have no effect on \( Y \) independently of \( X \).

(I4) I is (statistically) independent of any variable X that causes \( Y \) and that is on a directed path that does not go through \( X \). (2003: 98)

\(^{29}\) The notion of possibility that Woodward is working with here is the notion of logical possibility. Thus, “as long as there is some basis for assessing the truth of counterfactual claims concerning what would happen if various interventions were to occur, it doesn’t matter that it may not be physically impossible for those interventions to occur.” (2003: 130)
Thus, in its canonical formulation, interventionism is a non-reductive account of causation, whereby possible interventions are necessary components of causation, and interventions themselves are defined via causal notions. For ease of exposition I will be working with an abridged version of his account:

(M) X is a cause of Y with respect to variable set V iff there exists a possible surgical intervention on X with respect to Y such that all other variables in V are held fixed, and the value of Y changes.

An intervention is surgical “in the sense that it affects only X and what is caused by X and does not at the same time change other possibly confounding variables.” (Woodward 2008b: 138)

Interventions must be surgical so as to guard against mistaking mere correlation for genuine causation. For example, in order to determine whether a particular drug is instrumental to recovery from a particular disease, we see what happens when one group of patients ingest the drug (the study group) and another group of patients (the control group) does not ingest the drug, and patients are randomly assigned to the two groups in order to ensure that other possible causes (such as age) are held fixed. If ingesting or not ingesting the drug has a corresponding effect on recovery, we can say that the drug makes a difference to recovery.

2.3.2. Interventionist Solutions

2.3.2.1. The Interventionist Proportionality Argument

According to the interventionist proportionality argument, interventionism can be used to show how higher-level properties can be difference-makers insofar as they are proportionate to their effects in a way that their underlying lower-level properties are not. Consider the
following example: John wants a beer, believes that there is a beer in the refrigerator, and this belief causes him to go to the refrigerator.\textsuperscript{30} Now consider the following counterfactuals:

(1) If John’s belief that there is beer in the refrigerator were to be changed by an intervention to not having the belief, he would have gone to the grocery (and not to the refrigerator).

(2) If John’s brain state were to be changed by an intervention to not having that brain state, he would have gone to the grocery (and not to the refrigerator).

The idea here is that while (1) is true in the case where we change the belief from a belief that there is beer in the refrigerator ($X=x_1$) to a belief that there isn’t any beer in the refrigerator ($X=x_2$), it is plausible to say that (2) is false in a case where we change the underlying brain state from one of the possible realizing brain states of the belief that there is beer in the refrigerator ($B=b_1$) to a different possible realizing brain state of the belief that there is beer in the refrigerator ($B=b_2$). Thus, the belief that there is no beer in the refrigerator (rather than the belief that there is beer in the refrigerator) causes John to go to the grocery (and not to the refrigerator). However, it is (plausibly) not the case that if one of the realizing brain states of the belief that there beer in the refrigerator is changed to another possible realizing brain state of that belief, then John would go to the grocery (and not to the refrigerator).

Where does this leave \textit{Completeness}? According to the proportionality argument, \textit{Completeness} mistakes sufficient conditions for causes\textsuperscript{31}. This is important because on most

\textsuperscript{30} Discussion of this example is taken from Raatikainen (2010).

\textsuperscript{31} Proponents of interventionist solutions to the exclusion problem share this reading of the exclusion problem, and I will be assuming in what follows that they are right to do so. I should note, though, that one \textit{could} interpret the exclusion problem as dealing with \textit{causally} sufficient conditions, whatever notion of causation one has. In which case, it becomes less clear what the impact of incorporating the interventionist account of causation into the setup of the exclusion problem should be. In any case, I will be setting this issue aside for the purposes of this chapter.
interpretations of *Completeness*, the truth of *Completeness* only requires that the presence of P results in the presence of E, and yet the notion of a difference-making cause employed by interventionism *additionally* requires that the absence of P results in the absence of E. While both the belief that there is beer in the refrigerator and the realizing brain state of the belief are sufficient conditions for John’s going to the refrigerator, the absence of the belief that there is beer in the refrigerator results in the absence of John’s going to the refrigerator, whereas the absence of one realizing brain state of the belief plausibly still results in John’s going to the refrigerator. Thus, the belief that there is beer in the refrigerator is a difference-making cause of John’s going to the refrigerator, whereas the realizing brain state of the belief is *not* a difference-making cause of John’s going to the refrigerator.

Since the interventionist is interested specifically in difference-making causes, proponents of the proportionality argument suggest that the interventionist approach to the exclusion problem should replace “sufficient cause” with “difference-making cause” in the setup of the exclusion problem. In particular, Panu Raatikainen (2010: 361) suggests that *Completeness* should be revised as follows:

*Revised Completeness:* All physical occurrences that have difference-making causes have difference-making physical causes.

According to the proportionality argument, then, the exclusion problem is easily solved once we plug the interventionist account of causation into the setup of the problem. For once we do so, we see that *Revised Completeness* is plausibly false32: While there *is* a difference-

---

32 Presumably the rest of the setup of the exclusion problem would have to also be rewritten along difference-making terms in order to provide an adequate interventionist solution to the exclusion problem.
making cause of John’s going to the refrigerator, the difference-making cause needn’t be physical (or type-identical with a physical property). John’s belief that there is beer in the refrigerator is a difference-making cause of John’s going to the refrigerator, but the underlying brain state (as well as, presumably, any of the physical states underlying the brain state itself) is not a difference-making cause of John’s going to the refrigerator. As Peter Menzies puts it (in terms of M, rather than P, causing B):

So our epistemic situation is one in which we have to decide between the well-confirmed hypothesis that M is the cause of B and the purely conjectural hypothesis that there exists some physical property P that is a difference-making cause of B. It would not be irrational under these circumstances to favor the first hypothesis over the second, concluding that the strengthened causal closure principle [Revised Completeness] is false in this case. (Of course, consistently with thinking that M is the cause of B, we can still suppose that there exists a physical property that is causally sufficient for B. We have not found any reason to reject the causal closure of the physical [Completeness], as originally formulated. (Menzies 2008: 216)

2.3.2.2. The Interventionist Supervenience Argument

The supervenience argument challenges the overdetermination premises by drawing a distinction between standard cases of overdetermination (such as a death being caused by simultaneous shootings courtesy of distinct members of a firing squad) that result from two (or more) independent causes, and non-standard cases of overdetermination that result from two (or more) causes that bear some sort of metaphysical dependence relation to one another. There are two apparently different, but, according to Karen Bennett, actually equivalent, ways to state this challenge to the overdetermination premises:
(i) Non-standard cases of overdetermination differ from standard cases of overdetermination in that in the latter cases, the effect results from independent causes, whereas in the former cases, the effect results from dependent causes, i.e. causes that bear some sort of asymmetric metaphysical dependence relation to one another. We should apply the label “overdetermination” to the standard cases, and withhold the label from the non-standard cases. Thus, Overdetermination\textsubscript{1} is true but Overdetermination\textsubscript{2} is false.

(ii) Non-standard cases of overdetermination differ from standard cases of overdetermination in that in the latter cases, the effect results from independent causes, whereas in the former cases the effect results from dependent causes, i.e. causes that bear some sort of metaphysical dependence relation to one another. Both the latter and the former count as cases of overdetermination, but the latter are cases of “bad overdetermination” in the sense that it would be odd to think that most effects of mental causes take place in this manner, whereas the former are cases of “good overdetermination” in the sense that it would not be odd to think that most effects of mental causes take place in this manner. Thus, Overdetermination\textsubscript{2} is true but Overdetermination\textsubscript{1} is false.

According to Bennett (2003: 474), it is entirely a terminological issue whether we choose to distinguish between good cases of overdetermination (what takes place in cases of mental causation) versus bad cases of overdetermination (firing squad cases), or whether we choose to distinguish between cases that don’t involve overdetermination (mental causation cases) versus cases that do involve overdetermination (firing squad cases). What is important is the distinction between independent causes and dependent causes, and the interventionist can use this distinction to show how we can reject either Overdetermination\textsubscript{1} or Overdetermination\textsubscript{2}
(depending on whether we adopt (i) or (ii)) and still be able to say that there are cases of mental causation that are different from the standard cases of overdetermination.\textsuperscript{33}

A straightforward method\textsuperscript{34} for challenging the overdetermination premises is to (i) Provide a necessary condition for overdetermination; and (ii) Show that the firing squad cases meet the necessary condition for overdetermination, and also show that, given one’s favored dependence relation holding between the mental and the physical, there are some cases of mental causation that do not meet the necessary condition for overdetermination.

The interventionist version of a “very intuitive and widely accepted” (Bennett 2003: 476) necessary condition for overdetermination is as follows: \( E \) is overdetermined by \( M \) and \( P \) only if the following counterfactuals are both non-vacuously true:

\[
\text{[IO1]} \quad \text{If } P \text{ were varied via intervention while } M \text{ is held fixed, then } E \text{ would still have occurred;}
\]

and

\[
\text{[IO2]} \quad \text{If } M \text{ were varied via intervention while } P \text{ is held fixed, then } E \text{ would still have occurred.}
\]

The firing squad cases do indeed meet this necessary condition for overdetermination: If \( M \) is a shooting committed by one of the members of the firing squad, \( P \) is a shooting committed by a different member of the firing squad, and \( E \) is the death of the victim, then it looks as though both IO1 and IO2 are non-vacuously true: If we intervened on the shooting by the first

\textsuperscript{33} Hence my decision to label them Overdetermination1 and Overdetermination2 rather than Overdetermination and Exclusion.

\textsuperscript{34} The following is adapted from Bennett (2003).
member (say, by filling her gun with blanks) while holding the shooting by the second member fixed, the death would still have occurred. Likewise, if we intervened on the shooting by the second member while holding the shooting by the first member fixed, the death would still have occurred.

Woodward (2008a) and Lawrence Shapiro (2010; 2011) argue, in effect, that there are some cases of mental causation that do not meet the necessary condition for overdetermination given the impossibility (by hypothesis) of varying $M$ by means of an intervention while holding $P$ fixed. Now, nothing about the relationship between the mental and the physical follows from the interventionist account of causation per se. However, as discussed above, most nonreductive physicalists will agree that if mental properties supervene upon physical properties, then “it is metaphysically impossible for two subjects to differ with respect to their mental properties while sharing the same physical properties.” (Woodward 2008a: 256) Thus, “the common denominator” notion of supervenience shared by most non-reductive physicalists is the Supervenience relation articulated in 1.2.1, according to which the impossibility involved is stronger than mere nomological impossibility. So, if the interventionist qua non-reductive physicalist adopts Supervenience then, as Woodward points out, “[IO2] makes no sense or at least is inappropriate for capturing the causal influence of [M] on [E].” (2008a: 255) This is because if the supervenience relationship between the mental and the physical is stronger than mere nomological necessity, then it is impossible to vary $M$ by means of an intervention while holding $P$ fixed. It is impossible because according to Supervenience we can’t have an

35 That is, if we adopt Bennett’s preferred method of evaluating counterfactuals via deleting [a.k.a. using a “metaphysical hole-puncher” (2003: 482)], rather than via backtracking or replacing.
intervention on $M$ that changes $M$ while leaving $P$ unchanged, since that would involve a change in mental state with no corresponding change in the realizing neural state.

To illustrate the envisioned scenario, let’s fill in the details of this case using an example discussed by Woodward. Let $M$ stand for a sensorimotor intention\textsuperscript{36} to reach for the grape that is in front of the agent, $P$ stand for a certain pattern of neural firing in the parietal reach region of the posterior parietal cortex\textsuperscript{37}, and $E$ stand for the agent reaching for the grape. In order to evaluate $[IO2]$, it must be possible to vary the sensorimotor intention while holding the certain pattern of neural firing fixed. If, by hypothesis, sensorimotor intentions supervene on certain patterns of neural firing in the parietal reach region of the posterior parietal cortex, then it is impossible to vary the sensorimotor intention while holding the certain pattern of neural firing fixed.

Woodward and Shapiro have, in effect, identified an interventionist method for establishing a disanalogy between mental causation and the firing squad cases: In the firing squad cases, it is possible to perform the relevant intervention in order to determine the putative efficacy of the causes, and the necessary condition for overdetermination is met, since both (IO1) and (IO2) are non-vacuously true. Whereas in the mental causation cases, given the Supervenience relationship between the mental and the physical, it is not possible to perform the relevant intervention, and so the relevant counterfactual will be vacuously true, given that the antecedent (varying $M$ while holding $P$ fixed) will always be false. Thus, the firing squad

\textsuperscript{36} Here I am adopting Susan Pockett’s taxonomy, according to which willed intentions are “general plans to carry out a particular act”, and sensorimotor intentions “code for the relatively low-level and immediate motor plans.” (2006: 13)

\textsuperscript{37} See Andersen & Buneo (2002).
cases meet the necessary condition for overdetermination, whereas the mental causation cases do not.

2.4. Baumgartner’s Objection to the Interventionist Solutions

Baumgartner points out the following tension in the interventionist solutions (though he doesn’t distinguish between different versions of the interventionist solutions): While interventionists may succeed in challenging one (or more) of the exclusion problem premises, they are also unfortunately precluded from showing that there are any cases that meet the necessary conditions for mental properties to be causally efficacious. Baumgartner points out that the interventionist account of causation includes the following necessary condition for $X$ to be a cause of $Y$:

$Fix$: It must be possible to hold any variable in $V$ that isn’t located on a causal path from $X$ to $Y$ fixed, while the intervention is performed on $X$.\(^{38}\)

Yet, as we saw above, given the Supervenience relation between the mental and the physical it will be impossible to hold the physical subvenience base fixed while intervening on the supervening mental variable. Thus, the necessary condition for the causal efficacy of the mental cannot be met, and so interventionism cannot possibly vindicate the causal efficacy of

---

\(^{38}\) Baumgartner explicitly discusses $Fix$; Woodward does not.
the mental. I will now spell out how the interventionist proportionality and supervenience arguments (neither of which deny \textit{Fix}) both fall prey to Baumgartner’s objection.

The supervenience argument tries to use the fact that it is impossible to perform an intervention on M while holding P fixed to its advantage. According to Woodward, this means that “...the counterfactuals that give rise to the overdetermination present in the rifleman case \textit{lack any clear sense} in the cases in which there is a worry about the possibility of mental/physical overdetermination.” (2008a: 259, italics added) But Baumgartner points out that the interventionist is committed to saying something stronger here than simply pointing out that the counterfactual “lacks any clear sense” in the mental causation case. Rather, the interventionist is committed to saying that the necessary conditions for M to be a cause of E are \textit{not met} because, contra \textit{Fix}, there is no possible intervention such that, while it is performed on M, all variables in V that are not located on a causal path from M to E are held fixed. In general, when the necessary conditions for M to be a cause of E are not met this means that M does not cause E, and so the supervenience argument only works at the cost of giving up the claim that the mental is causally efficacious!

According to the proportionality argument, there are some interventions such that if John’s belief that there is beer in the refrigerator were to be changed by said intervention to not having the belief, he would have gone to the grocery (and not to the refrigerator). The problem, however, is that in order for such an intervention to be surgical, we would have to hold fixed the underlying brain state- which, given that the belief supervenes upon the underlying brain state, is impossible. Given that the variables representing John’s belief and the underlying brain state are both included in variable set V, we see that \textit{Fix} is violated. For there is a variable in V that isn’t located on a causal path from M to E, yet cannot be held fixed while the intervention is
performed on M. That variable is P (representing the brain state underlying the belief that there is beer in the refrigerator). Thus, a necessary condition for causation is not met.

Indeed, it turns out that rather than solving the exclusion problem, Baumgartner’s objection can be used to show that extant versions of interventionism are incapable of vindicating all higher-level causation, let alone mental causation. Indeed, this problem can be generated using only the assumptions of Supervenience and Fix: Take a higher-level variable X that supervenes upon lower level variable Z. According to Fix, if X is a cause of Y, then it must be possible to hold Z fixed while intervening on X. According to Supervenience, it is not possible to hold Z fixed while intervening on X. Therefore, X is not a cause of Y- and this conclusion holds for any higher-level variable that supervenes upon a lower-level variable. Ultimately I will provide a solution to this problem that (in part) relaxes the necessary condition on causation imposed by Fix. First, though, it is important to see why interventionists should not want to solve this problem by way of rejecting Supervenience.

To begin with, the considerations discussed in 2.2.1. apply not just to mental properties, but to a wide range of causally efficacious higher-level properties studied across the sciences whose interlevel relations seem best characterized by Supervenience, rather than identity or emergence. Furthermore, even if interventionists think that Supervenience is in fact false, this conclusion should, by the interventionists’ own lights, be drawn on the basis of empirical, rather than a priori, considerations. To put the point slightly differently, even if interventionists think that Supervenience is in fact false, this is different from the much stronger claim that Supervenience is impossible. Yet, if the solution to the problem above is to reject Supervenience, then the interventionist is, in effect, using her account of causation to rule out the very possibility of Supervenience!
On a related note, the final reason why interventionists should not want to solve the problem by way of rejecting *Supervenience* is that being *required* to reject supervenience in order to make sense of higher-level causation renders interventionism incapable of making *sense* of certain disputes in the sciences. There are *lots* of local instances of the dispute between identity, emergence and *Supervenience*, including

- the debate over whether classical genetics can be reduced to molecular biology;\(^{39}\)
- the debate over whether conscious properties are emergent\(^{40}\), type-identical with neural properties\(^{41}\), or realized by (but not type-identical with) neural properties;\(^{42}\)
- the debate discussed above concerning Chemical Emergentism.

A charitable interpretation of the disputes is that these are disputes amongst *coherent* options. Yet, if we adopt the strategy of rejecting *Supervenience*, interventionism is committed to saying that scientists who have embraced the middle ground between identity and emergence aren’t just wrong - their view is incoherent!

### 2.5 The Exemption Clause Strategy \([M^*]\)

Given the troubles that arise when trying to situate non-causal dependency relationships in standard causal graphs, an obvious solution presents itself. Namely, we should treat these non-causal dependency relationships *differently* from the relationships found in

---

\(^{39}\) See, e.g. Kitcher (1984); Waters (1990); Sarkar (1998); Rosenberg (2006).

\(^{40}\) E.g. Haskar (1999).

\(^{41}\) E.g. Hill (1991); Papineau (2002).

\(^{42}\) E.g. Lycan (1996); Dretske (1995); Tye (1995).
standard causal graphs- in particular, we should relax the interventionist necessary conditions on causation such that subvening bases *needn’t* be held fixed while the intervention on the supervening property is performed. Indeed, this is precisely the solution proposed by Woodward in response to Baumgartner.

Woodward (2014) suggests, in response to Baumgartner’s objection, that the interventionist account of causation be modified as follows:

(M*) X causes Y with respect to V iff there possibly exists a surgical* intervention on X with respect to Y such that all variables in V that are not related in terms of supervenience (or definition) to X or Y are held fixed, and the value of Y changes.43

This in turn leads to a revised version of Fix:

*Fix*: The variables in V that are not located on a causal path from X to Y and are not part of the subvenience base of X can be held fixed, while the intervention is performed on X.44

The idea, then, is that if the interventionist necessary conditions for causation specifically exclude requiring us to hold subvenience bases fixed while the intervention on the supervening property is performed, then we don’t have to worry about the necessary conditions

---

43 M* in the text is not directly taken from Woodward (2014). It is taken from Baumgartner (2013: 14), who explains in a footnote that “Woodward (2011) does not state M* explicitly, but merely indicates its relevant features. Thus, M* as given here, is my reconstruction of Woodward’s suggestion.” Woodward (2011) is the version of Woodward (2014) that was available on-line for several years prior to the publication of Woodward (2014). Also, for ease of exposition M* in the text is a slightly abridged version of M* in Baumgartner (2013). Instead of “surgical* intervention”, Baumgartner uses the phrase “(IV)*-defined intervention”, where the only difference between (IV)* and (IV) lies in the exemption clause for supervenience (and definition) relations. None of the other details of (IV)* are important for my purposes here, and so I have substituted “surgical intervention” for (IV)-defined intervention, and “surgical* intervention” for (IV)*-defined intervention.

44 This is taken from Baumgartner (2010: 378). Woodward himself doesn’t explicitly mention *Fix*. 
not being met in the compatibilist and incompatibilist approaches. Call this the exemption clause strategy [ECS].

Unfortunately, a new problem arises with the addition of the exemption clause. It turns out that while $M^*$ may be a necessary condition for $X$ causing $Y$, it cannot be a sufficient condition. To see how this problem arises, consider the following three graphs:

1st Possible Graph  2nd Possible Graph  3rd Possible Graph

Indicates a causal relation

Indicates a supervenience relation

The first possible graph depicts a scenario in which $X$ is a cause of $Y$, and $X$ supervenes upon $Z$, but $Z$ is not a cause of $Y$. The second possible graph depicts a scenario in which $X$ supervenes upon $Z$, and $Z$ is a cause of $Y$, but $X$ is not a cause of $Y$. The third possible graph depicts a scenario in which $X$ supervenes upon $Z$, and both $X$ and $Z$ are causes of $Y$. While one could, of course, insist that an account of causation should a priori rule out one or more of the above graphs, this would run counter to the empirical leanings of interventionism. One, if not the, core principle
underlying interventionism is the idea that we should let the empirical practices of scientists
guide us in our construction of a theory of causation. Moreover, we should be open to the
possibility of various causal relationships in the world being discovered. Thus, interventionism
should be open to the following three possibilities depicted in the graphs above:

- The possibility that a supervening property could be a difference-making cause of an
effect while its subvening base is not [1st possible graph];
- The possibility that a subvening property could be a difference-making cause of an
effect while its supervening property is not [2nd possible graph];
- The possibility that both a supervening property and its subvening base could be
difference-making causes of an effect [3rd possible graph]

Now, the problem with ECS is that, while it doesn’t rule out any of the three possibilities
above, it fails to distinguish between the three. I will first explain why a consequence of ECS is
the inability of interventionism to distinguish between the above three possibilities, and then I
will explain why this is a problem for interventionism.

According to ECS, it needn’t be possible to hold X’s subvenience base Z fixed in order to
perform a surgical intervention on X. The first thing to notice is that this results in an epistemic
situation whereby the results of the surgical* intervention would give us no reason to prefer one
of the three possible graphs over the others. Remember, according to ECS, if a surgical*
intervention on X results in a difference in Y, then X is a difference-making cause of Y. However,
if a surgical* intervention on X results in a difference in Y, this actually gives us no reason to
think that X, rather than Z, is a difference-making cause of Y. This is because a surgical*
intervention on X resulting in a difference in Y is perfectly compatible with the following two scenarios:

Scenario 1

Scenario 2

Given that X supervenes upon Z, any intervention on X will also be an intervention on Z—this is just another way of saying that it is impossible to intervene on X while holding Z fixed.

Scenario 1 is a case in which X is a difference making cause of Y while Z is not, and Scenario 2 is a case in which Z is a difference making cause of Y while X is not. Now, if all we know is that a surgical* intervention on X results in a difference in Y, this could be because X is a difference making cause of Y (Scenario 1) or it could be because Z is a difference making cause of Y (Scenario 2). Of course, in a situation that doesn’t involve non-causal dependency relationships, this problem doesn’t arise, for we could, and would, simply hold Z fixed while intervening on X. But since X supervenes upon Z we can’t hold Z fixed while intervening on X, and therefore if a surgical* intervention on X results in a difference in Y, this gives us no reason to think that X, rather than Z, is a difference-making cause of Y.

The reasoning above shows that one consequence of ECS is that it results in an epistemic situation whereby the results of a surgical* intervention on X give us no reason to prefer the first possible graph over the second possible graph. If that’s true, then it is also the
case that the results of a surgical* intervention on X cannot be used to prefer the third possible graph over either the first or second possible graphs. According to the third possible graph, both X and Z are difference-making causes of Y. However, if both X and Z are difference-making causes of Y, it of course must be the case that X is a difference-making cause of Y. As we already saw, the results of a surgical* intervention on X cannot be used to show that X is a difference-making cause of Y, for the results can equally justify the inference that Z is a difference-making cause of Y. Thus, if a surgical* intervention on X makes a change to Y, this gives us no reason to prefer any of the three possible graphs: It may be the case that X (but not Z) is a difference-making cause; it may be the case that Z (but not X) is a difference-making cause; and it may be the case that both X and Z are difference-making causes.

So far we have been focusing on the epistemic limitations that result from ECS. However, the consequences of ECS extend beyond merely epistemic limitations. For on the interventionist picture, it is the epistemology that drives the metaphysics: What it is for X to cause Y turns out to be nothing more than what it would take for X to satisfy the (ideal) conditions imposed by our best experimental techniques aimed at causal inference. So the problem is not (just) that it is impossible to use interventionism to experimentally distinguish between the three possible graphs. The problem is that, if ECS is the right way to deal with non-causal dependency relationships, then interventionism is forced to say that there is no difference between

- What it is for a supervening property to be a difference-making cause of an effect while its subvening base is not;
• What it is for a subvening property to be a difference-making cause of an effect while its supervening property is not;

• What it is for both a supervening property and its subvening property to be difference-making causes of an effect

This is a problem because, at least intuitively, it seems as though the three possible graphs are describing three distinct possibilities: A supervening property being a difference-making cause of an effect is not the same thing as its subvening property being a difference-making cause of the effect, and vice-versa. And it is not the same thing for both a supervening property and its subvening base to be difference-making causes of an effect as it is for the supervening (but not subvening) property to be a difference-making cause of the effect, or for the subvening (but not supervening) property to be a difference-making cause of the effect. Thus, while $M^*$ may be a necessary condition for $X$ being a cause of $Y$, it can’t be sufficient, given that $M^*$ is consistent with $Z$ rather than $X$ being a cause of $Y$.

One possible response on behalf of interventionism is to simply insist that if a surgical intervention on $X$ results in a difference in $Y$, then it must be the case that $X$, rather than $Z$, is a difference-maker. In effect, this response involves ruling out the very possibility that $X$ could be merely correlated with $Y$, while $Z$, rather than $X$, is a genuine cause of $Y$. However, as noted above, the empirical leanings of interventionism would seem to run counter to this possible response. For interventionism should at least be open to the possibility that $X$ could be merely correlated with $Y$, and it is then (says the interventionist) simply a matter of running the right experiments and seeing whether or not there are in fact cases where $X$ is merely correlated with $Y$ while $Z$, rather than $X$, is a genuine cause of $Y$. 

43
2.6. The Multiple Intervention Strategy \([M^{**}]\)

A second possible response on behalf of interventionism is to switch from the single-intervention approach of \(M\) and \(M^*\) to a multiple-intervention approach. The idea here is that perhaps what has gone wrong in \(M\) and \(M^*\) is that we have been trying to use the results of interventions on a single variable (X) in order to determine that X is a cause of Y, when in fact we need to use the results of interventions on multiple variables (X and Z) in order to determine that X is a cause of Y. Here is the proposal in more detail:

\((M^{**})\) X is a cause of Y with respect to \(V\) iff

(a) there possibly exists a surgical* intervention on X with respect to Y such that all variables in \(V\) that are not related in terms of supervenience (or definition) to X or Y are held fixed, and the value of Y changes;[this first conjunct is just \(M^*\)] and

(b) there possibly exists a surgical* intervention on any variable in \(V\) that is a subvenience base such that all other variables (including X) are held fixed and the value of Y does not change\(^{45}\)

This proposal- which makes no changes to \(\text{Fix}^*\)- is an improvement insofar as it enables us to distinguish between the first possible graph and the second possible graph on the basis of using the results of interventions on multiple variables (X and Z) in order to determine that X, rather than Z, is a cause of Y. Consider, again, the case of John and his quest for beer, and consider the following possible values for variables X, Z, and Y.\(^{46}\)

\(^{45}\) Unlike \(M\) and \(M^*\), \(M^{**}\) is a proposal original to me, not taken from either Woodward or Baumgartner.

\(^{46}\) In listing the various possible values, or realizations, I am trying to steer clear of disputes about the nature of realization and multiple realization. I do, however, think that my discussion more naturally fits
As we saw earlier, it is at least plausible that there will be cases where no possible surgical intervention on Z results in a change in Y. Starting with Z1: It is plausible that if we hold X fixed (at X1) while varying Z1 via intervention (the situation in which this is possible is one where (Z2 is activated when Z1 is not activated), there will be no resulting change in Y. That is to say, the individual’s intention to go to the refrigerator will cause her to go to the refrigerator regardless of whether Z = Z1 or whether instead Z = Z2. And the same (again, plausibly) goes for holding X fixed at X2: The individual’s intention to go to the grocery will cause her to go to the grocery regardless of whether Z = Z3 or whether Z = Z4.

It is also plausible that there will be cases where even though no possible surgical intervention on Z results in a change in Y, surgical* interventions on Z (i.e. where X is not held with Gillett’s (2002) dimensioned view- which, unlike the flat accounts implicitly endorsed by other interventionists, is indeed compatible with massive multiple realization and thereby (perhaps) does not fall prey to Shapiro’s (2011) objection. For more discussion see, inter alia, Aizawa & Gillett (2009a; b); Polger (2008); Polger & Shapiro (2008).
fixed) will result in a change in Y. That is to say, if when we vary Z1 we also change the intention—
either from an intention to go to the refrigerator to an intention to go to the grocery, or vice
versa—it is also plausible that this will have a corresponding effect on whether the individual
goes to the refrigerator or whether she goes to the grocery.

Thus, if a surgical* intervention on Z results in a change in Y, but a surgical intervention
on Z does not result in a change in Y, we conclude that Z is not a cause of Y. Of course, from an
epistemic standpoint it is more difficult to use the interventionist framework to show that a
given variable is not a cause— which requires ruling out the possibility of any surgical intervention
on it resulting in a corresponding effect— than to show that a given variable is a cause, which
merely requires one possible surgical intervention. I am assuming that the case above, and
others like it, are cases where it is at least plausible to think that there will be no possible
surgical interventions on Z that result in Y, though of course in practice it may be difficult to rule
out the possibility of any surgical intervention on Z resulting in a change in Y. The point here
though is metaphysical, not epistemic: In a case where there is no possible surgical intervention
on Z that results in a change in Y, Z is not a cause of Y.

The interventionist can now distinguish the first possible graph from the second possible
graph on the basis of M**: According to M**, a surgical* intervention on X resulting in a change
in Y is necessary, but not sufficient, for X to be a cause of Y. In order for X to be a cause of Y, it
must additionally be the case that a surgical* intervention on Z does not result in a change in Y.
Thus, the way to distinguish scenarios 1 and 2 above is to additionally perform a surgical*
intervention on Z while holding X fixed. Scenario 1 is a case where a surgical* intervention on X
results in a change in Y and a surgical* intervention on Z does not result in a change in Y.
Whereas in Scenario 2, a surgical* intervention on Z does result in a change in Y. Thus, in
contrast with \( M^* \), \( M^{**} \) provides a necessary \textit{and} sufficient condition for \( X \) for being a cause of \( Y \), and it is \textit{not} consistent with \( Z \) rather than \( X \) being a cause of \( Y \).

Of course, in Scenario 2, if \( Z \) supervenes on \( Z-1 \), then we would have to additionally use the results of a surgical* intervention on \( Z-1 \) in order to determine whether \( Z \) is a cause of \( Y \).

Thus, \( M^{**} \) has an iterative process built in: We use surgical* interventions on both the supervening and subvening properties in order to determine whether the supervening property is a difference-making cause. But if the results of these surgical* interventions do not rule in favor of the supervening property- that is, it turns out that a surgical* intervention on \( X \) results in a change in \( Y \) and a surgical* intervention on \( Z \) results in a change in \( Y \) as well- we must then apply the same process to \( Z \) and its subvenience base \( Z-1 \) in order to determine whether \( Z \) is a cause of \( Y \). There is, however, no \textit{problematic} regress here: We don’t need to keep performing surgical* interventions at lower and lower levels, \textit{ad infinitum}, in order to determine that \( X \) is a cause of \( Y \). In order to determine that \( X \) is a cause of \( Y \), all we need are surgical* interventions on \( X \) and \( Z \). And if \( X \) \textit{isn’t} a cause of \( Y \), in order to determine that \( Z \) is a cause of \( Y \) all we need are surgical* interventions on \( Z \) and \( Z-1 \).

\textbf{2.7.} \( M^{**} \) and Supervenient Overdetermination

While \( M^{**} \) is an improvement over its predecessors, it does not fully solve the problem we began with when evaluating \( M^* \). For although \( M^{**} \) enables interventionism to distinguish between the first and second possible graphs, it does so at the cost of ruling out the possibility of the third possible graph. According to \( M^{**} \), the only situation in which we can say that \( X \) is a cause of \( Y \) is the situation where \( Z \) has been ruled out as a cause of \( Y \). If that’s the case, then
according to **M** there can’t be a situation where both X and Z are causes of Y, since X being a cause of Y precludes the possibility of Z being a cause of Y: X is a cause of Y only if Z is not a cause of Y. Thus, if **M** is the right way to deal with non-causal dependency relationships, then interventionism is forced to say that it is impossible for both a supervening property and its subvening base to be causes of the same effect. Call this the impossibility of supervenient overdetermination.

If **M** were the only version of interventionism that provides us with necessary and sufficient conditions for higher-level causation, then the fact that it rules out the possibility of supervenient overdetermination could, in theory, be used to argue against interventionism: Given that the possibility of supervenient overdetermination seems like a genuine possibility, it is therefore a mark against interventionism that it is forced to rule out the possibility of what seems like a genuinely possible scenario.

However, even if ruling out the possibility of supervenient overdetermination should be seen as a cost of **M**- I shall question this assumption below- it seems like a relatively small cost that needn’t result in the wholesale rejection of interventionism. It is important to point out that there are plenty of cases of overdetermination whose possibility is not ruled out by **M**. For example, the possibility of the standard cases of overdetermination involving firing squads is not ruled out by **M**, as these cases do not involve supervenience relations. Nor is the possibility of overdetermining emergent causes ruled out by **M**, for emergent properties and their bases- that is, what they emerge from- can, in principle, be held fixed while the other is manipulated. **M** only rules out the possibility of overdetermination in cases where one cannot, in principle, hold the subvening variable fixed while manipulating the supervening variable.
That said, I don’t think it is clear that ruling out the possibility of supervenient overdetermination is a cost at all. Instead, we could view this as an unexpected benefit of the interventionist analysis of causation: Interventionism gives us deep insight into what is actually possible. While, prior to interventionism, we may have thought that cases of supervenient overdetermination are at least possible, interventionism shows us that they are not, in fact, possible. Indeed, the fact that $M^{**}$ precludes the possibility of supervenient overdetermination would only be problematic if we have a strong reason to accept at face value the appearance of supervenient overdetermination being genuinely possible. However, if the appearance of supervenient overdetermination being genuinely possible can be explained away as resulting from mistaken intuitions, then it looks as though ruling out the possibility of supervenient overdetermination shouldn’t even be seen as a small cost of interventionism. More empirical research would of course need to be done to fully explain why the appearance of supervenient overdetermination being genuinely possible exists, and where it comes from. Here, though, are some suggestive explanations for why supervenient overdetermination appears to be genuinely possible, even though $M^{**}$ shows us that it isn’t.

First, it may be the case that when it seems to us as though supervenient overdetermination is genuinely possible, we are simply transferring intuitions that we have about standard cases of overdetermination to cases of supervenient overdetermination while neglecting to track the switch from non-supervening variables in standard cases of overdetermination to supervening variables in cases of supervenient overdetermination. Thus, given that overdetermination in firing squad cases appears to be genuinely possible, perhaps we simply transfer that intuition to cases of supervenient overdetermination.
Second, it may be the case that when it seems to us as though a theory of causation oughtn’t rule out the possibility of supervenient overdetermination, we are in fact relying upon theories of causation that may allow for the possibility of supervenient overdetermination but stand in opposition to the interventionist account of causation. For example, on a nomic subsumption theory of causation (the theory of causation that seems to be in the background of the exclusion problem), there doesn’t seem to be anything in the theory itself that precludes the possibility of supervenient overdetermination. The theory doesn’t preclude the possibility of their being multiple sufficient conditions at the same time, and there is no reason to think that the possibility is precluded simply in virtue of the conditions bearing a supervenience relation to one another.

In general, then, given that it is not at all clear where our intuitions about causes have been coming from in cases involving supervenient overdetermination, we should be open to the possibility that these intuitions are simply mistaken, especially in light of the theoretically motivated preclusion of the possibility of supervenient overdetermination provided by M**. Before moving on and using M** to revisit the exclusion problem, it will be useful to provide a recap of the different options, and associated costs, available to interventionism in light of Baumgartner’s objection:

- M is only capable of vindicating the causal efficacy of higher-level properties at the cost of ruling out the very possibility of *Supervenience*
- M* is only capable of vindicating the causal efficacy of higher-level properties at the cost of ruling out the very possibility that X could be merely correlated with Y, while Z, rather than X, is a genuine cause of Y
* M** is only capable of vindicating the causal efficacy of higher-level properties at the cost of ruling out the very possibility of supervenient overdetermination.

Notice, first, that the costs of all three versions of interventionism can be side-stepped by rejecting *Higher Level Efficacy*: Higher-level properties studied across the sciences can be causes. Although we already saw in 2.2.2. why there is good reason to not reject *Efficacy*, it is important to see why interventionists should also not reject *Higher Level Efficacy*: If *Higher Level Efficacy* is false, then either causation only *really* takes place at the bottom level of physics\(^{47}\), or causation *never* actually takes place, even in physics.\(^{48}\) However, if either of these positions is true, then this would mean that *all* of the causal talk in the special science is simply false. And if that’s the case, then interventionism would find itself in the odd position of having constructed an account of causation that is based on what investigators in the special sciences seem to be looking for when constructing and conducting experiments and models used to make distinctively *causal* inferences, even though, if *Higher Level Efficacy* is false, then all of the causal inferences made in the special sciences are false as well. This would be akin to constructing an account of causation based on the practices of astrologers, even though one knows that all of the causal inferences made in astrology are false. Indeed, Woodward himself advertises interventionism as being uniquely capable of making sense of scientific practices *outside of physics* that involve the notion of causation. He writes:

This ‘manipulationist’ [interventionist] conception of causal explanation has the advantage of fitting a wide range of scientific contexts, especially in the social and behavioral sciences, where investigators think of themselves as discovering casual relationships and constructing causal explanations, but where

\(^{47}\) See, e.g. Salmon (1984); Dowe (2000).

\(^{48}\) For a good collection of articles on this topic, see Price & Corry (2007).
narrower notions of causal explanation, such as Salmon’s, seem to be of very limited applicability.

(Woodward 2003: 6-7)

Now, back to comparing and contrasting $M, M^*, \text{ and } M^{**}$: I have argued (on behalf of $M^{**}$) that if having to rule out the very possibility of supervenient overdetermination is a cost, it is a small cost, and it pales in comparison to the costs associated with $M$ and $M^*$. Furthermore, unlike the possibilities ruled out by $M$ and $M^*$, I think it is plausible that the “possibility” ruled out by $M^{**}$ is simply an artifact of mistaken intuitions easily explained away, and so the “cost” associated with adopting $M^{**}$ is in fact a strength: $M^{**}$ vindicates the causal efficacy of higher-level properties and provides insight into the nature of overdetermination.

2.8. Interventionism and the Exclusion Problem Revisited

I have argued that $M^{**}$ (a) is a plausible version of interventionism that can provide necessary and sufficient conditions for higher-level causation; and (b) it rules out the possibility of supervenient overdetermination. (a) provides ammunition on behalf of the proportionality and supervenience arguments in response to Baumgartner’s objection. However, (b) shows us that the supervenience argument cannot be sound. For according to the supervenience argument, *pace* the overdetermination premises both $M$ and $P$ can be causes of $E$ when $M$ supervenes upon $P$. Yet, according to $M^{**}$, it is impossible for both $M$ and $P$ to be causes of $E$ when $M$ supervenes upon $P$. 
Indeed, in light of M**, Overdetermination1 should be made stronger.\textsuperscript{49}

Overdetermination1 allows for the possibility of some instances in which E is overdetermined by both M and P; Overdetermination1 simply insists that at least some cases of mental causation don’t involve (supervenient) overdetermination. But given the impossibility of supervenient overdetermination, Overdetermination1 should insist that there are no instances in which E is overdetermined by both M and P. M** also provides a new gloss on Bennett’s ‘terminological issue’ approach to overdetermination. Remember, according to Bennett, it is entirely a terminological issue whether we choose to distinguish between good cases of overdetermination (what takes place in cases of mental causation) versus bad cases of overdetermination (firing squad cases), or whether we choose to distinguish between cases that don’t involve overdetermination (mental causation cases) versus cases that do involve overdetermination (firing squad cases). However, in light of M**, the contrast is really between possible cases of overdetermination (firing squad cases) and impossible cases of overdetermination (mental causation cases).

There is, then, an interesting relationship between interventionism and the exclusion problem: Reflection on the exclusion problem reveals inadequacies with extant versions of interventionism, and has led us to a new version of interventionism (M**) that, in addition to vindicating the causal efficacy of higher-level properties, also provides insight into the nature of overdetermination- namely, the appearances of supervenient overdetermination are deceiving, for supervenient overdetermination is actually impossible. This insight, in turn, reveals that

\textsuperscript{49} Of course, the interventionist would also have to rewrite Overdetermination along difference-making terms. See above, fn. 31.
Overdetermination must be strengthened, and that the Supervenience Argument approach to the exclusion problem cannot be sound.

Thus, at least given the considerations discussed in this chapter, it turns out that the only viable interventionist solution to the exclusion problem on behalf of non-reductive physicalism is provided by the interventionist proportionality argument. The interventionist proportionality argument does not insist that both M and P can be causes of E when M supervenes upon P, and therefore it does not run afoul of the impossibility of supervenient overdetermination. And while the interventionist proportionality argument cannot meet a necessary condition for causation according to M, it can meet the necessary conditions for causation according to M**, for according to M** we needn’t hold fixed the underlying brain state in order for a belief to be a difference-making cause. I conclude that Interventionism can be used to solve the exclusion problem, but only by making use of my M** version of interventionism.
Chapter 3: Making Sense of the Different Senses of Explanation

Philosophers of science have often alluded to two different senses of explanation\(^{50}\), as illustrated in the following sentences:

(1) The force of gravity acting on the apple is the explanation for why the apple fell to the ground.

(2) A standard physics textbook contains the explanation for why the apple fell to the ground.

In (1), the term ‘explanation’ refers to a causal force ‘out there in the world’, whereas in (2) ‘explanation’ refers to a description, or representation, of the causal force. Less commonly, philosophers of science have sometimes argued that these two different senses of explanation are in *conflict* with one another. Indeed, Wesley Salmon (1984; 1989) distinguishes theories of explanation in terms of which sense of explanation one’s theory of explanation emphasizes:

\(^{50}\) See, *inter alia*, Lewis (1986:217-18); Bird (2005); Salmon (1984; 1989); Craver (2007; 2013); Bechtel (2008); Waskan (2006: 201); Strevens (2008: 6); Illari & Williamson (2010: 280); Wright (2012); Illari (2013). For ease of exposition I begin my discussion with these two senses of explanation, though later in the paper I disambiguate ‘explanation’ more finely into three senses rather than two.
Ontic construals of explanation emphasize the sense of explanation used in (1), whereas epistemic construals of explanation emphasize the sense of explanation used in (2).51

The most recent version of the debate as to which sense of explanation should be emphasized takes place within the context of an internecine debate amongst the New Mechanists—recent proponents of a mechanistic approach to scientific explanation. Thus, some New Mechanists (such as Carl Craver) argue that mechanistic explanations are best construed ontically [O-theorists], while others (such as William Bechtel) argue that mechanistic explanations are best construed epistemically [E-theorists].

In this chapter I show that no good arguments have been offered in support of the claim that the different senses of explanation are in conflict with one another. Indeed, I argue that the debate at hand is deeply confused insofar as it concerns a pseudo problem—one that arises from a failure to recognize a widely appreciated claim about thoughts: Namely, that in many circumstances thoughts are individuated by their referential contents, which in turn needn’t require cognizers at all.

In 3.1 I sketch the origins of the EVO debate so as to distinguish the debate in its current guise from the debate in its original form. Participants in the debate typically reject the claim that explanation has multiple, legitimate senses, and so in 3.2. I present prima facie evidence in favor of ‘explanation’ being a polysemous term—that is, a term with multiple (though related) legitimate senses. In 3.3. I present the debate in its current guise between Craver and Bechtel, and I propose that a charitable interpretation of the dispute reveals two quite different possible disagreements: The first, which I shall refer to as the Uniqueness Dispute, presupposes that

51 Bird (2005) uses the term ‘objectivist’ and ‘subjectivist’ instead of ‘ontic’ and ‘epistemic’.
there is one and only one thing uniquely worthy of being called ‘explanation’, and the dispute then is over what the correct referent of ‘explanation’ actually is. The second, which I shall refer to as the Fundamentality Dispute, allows for more than one legitimate sense of explanation, but insists that one of the senses of explanation is more fundamental than the other sense(s). The dispute, then, is over which of the senses of explanation is the fundamental sense of explanation. In 3.4. I reject the main assumption underlying the Uniqueness Dispute by providing additional evidence in favor of ‘explanation’ being a polysemous term, and in 3.5. I argue that there is no good reason to think that one of the senses of explanation is more fundamental than the other sense(s). Finally, in 3.6. I show why the debate is deeply confused.

3.1. Origins of the EVO Debate

The first step towards removing the confusions involved in the EVO debate is to distinguish the debate in its current guise - a debate amongst proponents of one approach to scientific explanation, i.e. the New Mechanists - from the debate in its original form, which was a debate between proponents of different approaches to scientific explanation. The ontic/epistemic distinction was originally drawn by Wesley Salmon, whose causal-mechanical model of explanation was the first fully developed, and self-professed, version of an ontic model of explanation. According to Salmon’s causal-mechanical model of explanation, scientific explanation consists in showing why an event occurred by identifying the causal process(es) responsible for the occurrence of the explanandum. Often this will involve identifying “the
hidden mechanisms by which nature works”

Salmon changed his mind over the years with respect to which account of causation is best-suited for the explanatory role that causation plays in the causal-mechanical model.

This is one of the contributions that the New Mechanists make to the debate.

Hempel (1965); Van Fraasen (1980); Kitcher (1989).
As originally formulated by Salmon, the ontic/epistemic distinction pointed towards an incompatible relation between O-theorists and E-theorists, stemming (in part) from the epistemic relativization objection to Hempel’s attempted extension of the DN model to statistical explanations. The epistemic relativization objection arose when, in response to the reference class problem associated with inductive explanations, Hempel proposed the requirement of maximal specificity, which aims to restrict the evidence that can be cited in the inductive explanations to a specific body of knowledge at the time in question. The example that Hempel discusses involves John Jones recovering from a strep infection. On the one hand, John Jones’s strep infection may belong to the reference class of strep infections that are resistant to penicillin, in which case if he is given penicillin than his recovery from the strep infection is not highly probable. The requirement of maximal specificity tells us to assign the explanandum event to the most specific reference class to which it is known to belong. Thus, in a case where we know that John Jones’s strep infection is resistant to penicillin, we cannot simply assign it to the reference class of strep infections in general (which, when treated with penicillin, have a high probability of recovery). Rather, we must assign it to the reference class of strep infections that are resistant to penicillin.

The objection raised (by Salmon and others) to the requirement of maximal specificity is that it makes the goodness of inductive explanations relative to what scientists happen to know at a given time. If the strep infection is resistant to penicillin but we don’t know that it is, then on Hempel’s account we can give a good explanation of Jones’s recovery from the infection by citing the probabilistic law involving the recovery from strep infections when given penicillin, and the fact that Jones was given penicillin. Yet, if the infection is in fact resistant to penicillin, the explanation may be good relative to the explainer’s epistemic situation, but it is not a good
explanation if our evaluative standards involve objective correctness. Thus, Salmon argues that
if Jones’s infection is in fact resistant to penicillin, then the above explanation is not a good
explanation, for if it were, then the goodness of explanations would be relative to particular
epistemic situations at the expense of being relative to states of affairs actually obtaining in the
world.

This objection to the requirement of maximal specificity was initially used by Salmon to
underscore a fundamental difference between E-theorists and O-theorists: E-theorists
(according to Salmon) are willing to grant that the goodness of explanations can be relative to
particular epistemic situations at the expense of being relative to states of affairs actually
obtaining in the world, whereas O-theorists insist that the goodness of explanations should
not be relative to particular epistemic situations at the expense of being relative to states of
affairs actually obtaining in the world. On this understanding, then, the EVO debate in its
original form was simply a localized version of scientific realism versus instrumentalism, with the
O-theorists siding with scientific realism and the E-theorists siding with instrumentalism.

Moreover, in later years Salmon appeared to grant that he had no quarrel with E-
theorists who endorsed scientific realism, and therefore he thought it was perfectly harmless for
O-theorists such as himself to speak of “explanation” in both an ontic and an epistemic sense:
Proponents of this [ontic] conception can speak in either of two ways about the relationship between
explanations and the world. First, one can say that explanations exist in the world. The explanation of
some fact is whatever produced it or brought it about. The explanation consists of certain particular facts
and lawful relationships. The explanandum is also some fact. This manner of speaking will sound strange
to philosophers who have been strongly influenced by the thesis that explanations are arguments, or by
the deeply linguistic approaches that regard explanations as speech acts. In nonphilosophical contexts,
however, it seems entirely appropriate to say such things as that the gravitational attraction of the moon explains the tides, or the drop in temperature explains the bursting of the pipes. The gravitational attraction and the drop in temperature are out there in the physical world; they are neither linguistic entities (sentences) nor abstract entities (propositions). Second, the advocate of the ontic conception can say that an explanation is something—consisting of sentences or propositions—that reports such facts. It seems to me that either way of putting the ontic conception is acceptable; one can properly say either that the explanandum-fact is explained by the explanans-facts or that the explanans-statements explain the explanandum-statement. (Salmon 1989: 86)

At this point there is a strong temptation to take a page from Carnap's *Logical Foundations of Probability* and announce that there are two concepts of scientific explanation—explanation$_1$ and explanation$_2$—both of which are perfectly legitimate, and which must not be confused with one another...If the foregoing suggestions are correct, we can reconcile the currently viable versions of the epistemic and ontic conceptions. (Salmon 1989: 184)

Thus, the original version of the EVO debate was more or less resolved by 1989, on the assumption that neither O-theorists nor E-theorists are *committed* to denying an at least minimal version of scientific realism. Moreover, the resolution of the debate presupposed the existence of (at least) two, *equally legitimate*, senses of explanation. In its current guise, it is implausible to see the EVO debate as a localized version of scientific realism versus instrumentalism, for the E-theorists participating in the (current) EVO debate are not opponents of scientific realism. For example, in dismissing the ontic construal of mechanistic explanation as a trivial addition to what is already accepted by proponents of mechanistic explanations, Wright & Bechtel write: “Do the component parts and their operations and organization figure in our understanding of how and why depolarization occurs? Well, yes, in a flat-footed sense: without
any of these things to implicate, mechanistic explanations would be without content.” In addition, the positive elements of the epistemic construal that E-theorists emphasize don’t focus on taking an instrumentalist stance.

Though it is far from clear, then, what exactly is at issue in the current guise of the EVO debate, participants in the debate seem to have rejected the claim that the term ‘explanation’ has multiple, legitimate, senses. In the next section I present *prima facie* evidence in favor of ‘explanation’ being a polysemous term - that is, a term with multiple (though related) legitimate senses.

3.2. The *prima facie* case for ‘explanation’ being a polysemous term

In this section I briefly present the *prima facie* evidence in favor of ‘explanation’ being a polysemous term. In many contexts it is harmless to use the term ‘explanation’ without clarifying which sense of the term one is using. However, in the context of evaluating the EVO debate it is critical to distinguish between the different senses discussed herein.

3.2.1. The Representation/World Ambiguity

Explaining acts produce something with representational content. When we are using the term ‘explanation’ to refer to the content of the explanatory product, we may be referring either to the representations (public or private) individuated by the content, or we may be referring to states of affairs ‘out there in the world’ that either are, or *could* be, picked out by a

---

57 Wright & Bechtel 2007: 50.
representation- even if the modality here involves only a metaphysical, or logical, possibility.\textsuperscript{58}

Consider the following example from neurophysiology:

(3) The temporally coordinated opening and closing of transmembrane channels are the explanation for the conductance changes in action potentials.

In this example, we can use the term ‘explanation’ to refer to the representations of the relationship between conductance changes in action potentials and the temporally coordinated opening and closing of transmembrane channels as found in a standard neuroscience textbook. We can also use the term ‘explanation’ to refer to the temporally coordinated opening and closing of transmembrane channels themselves; moreover, these needn’t be picked out by an actual representation. Thus, in the 18\textsuperscript{th} century there were no actual explaining acts whose products referred to the conductance changes in action potentials, the temporally coordinated opening and closing of transmembrane channels, and the causal relationship between the two. Nonetheless, assuming that the action potentials that occurred in the 18\textsuperscript{th} century worked in exactly the same way as they do in the 21\textsuperscript{st} century, we can say that in the 18\textsuperscript{th} century, with respect to action potentials, there existed explanatory states of affairs that could be picked out by the product of an explaining act.

\textbf{3.2.2. The Process (Act)/Product Ambiguity}\textsuperscript{59}

\footnote{\textsuperscript{58} I am using the term ‘states of affairs’ quite broadly to include all sorts of mind-independent states, processes, properties, entities, etc.}
In its verb form, the term ‘explanation’ can be used to refer either to an act of explaining, or to the product of an explaining act. Consider the following example:

(4) The neuroscientist provided us with the explanation for the conductance changes in action potentials.

We can use the term ‘explanation’ in this case to refer to a particular individual engaged in a particular explaining act- that is, the individual verbally explains the conductance changes in action potentials to an audience by describing the temporally coordinated opening and closing of transmembrane channels. We can also use the term ‘explanation’ in this case to refer to the product of the individual’s explaining, for explaining acts produce something (e.g. words, sentences, images, etc.). Thus, the conductance changes in action potentials may be explained by a textbook description of the temporally coordinated opening and closing of transmembrane channels.

The ambiguities discussed in this section point to the following three senses of ‘explanation’ that capture different possible referents of ‘explanation’:

- Communicative Sense- a particular subset of communicative acts

---

59 This ambiguity has been noted by many explanation theorists, including Bromberger (1962), Achinstein (1983), Lewis (1986), Ruben (1990), and Salmon (1992).

60 As Ruben points out, ‘explanation’ isn’t the only term to suffer from this kind of ambiguity. Other examples that Ruben lists include “philosophically uninteresting ones like ‘simulation’ and destruction’”, as well as those that “raise philosophical issues similar to those raised by ‘explanation’: ‘prediction’, ‘derivation’, ‘proposition’, ‘argument’, ‘statement’, and ‘analysis’.” (Ruben 1990: 7)
● Epistemic Sense- a particular subset of representations
● Ontic Sense- a particular subset of states of affairs

The epistemic and ontic senses of explanation will be the primary focus of what follows, and it is important to stress that from the perspective of trying to make sense of scientific practice, both the epistemic and ontic senses at least appear to figure prominently in scientific practice: When we compare the relative merits of different explanations for the same explanandum, for example- a practice ubiquitous across the sciences- we are (typically) comparing the relative merits of different representations of the way things in the world work. And unless one adopts an extreme anti-realist interpretation of scientific practice (which none of the participants in the EVO debate profess to do), the referents of the explanatory representations are supposed to be real states of affairs. Thus, from the perspective of trying to make sense of scientific practice, there is prima facie evidence in favor of multiple senses of the term ‘explanation’.

3.3. The Debate Between Craver and Bechtel

According to mechanistic accounts of scientific explanation, scientific explanation (at least in many of the life sciences) consists in showing how the phenomenon in question is produced by a mechanism. While the New Mechanists share much in common, one area of disagreement concerns whether the mechanistic account is best construed ontically (Craver) or epistemically (Bechtel). According to the ontic construal of mechanistic explanation,
“explanations are objective features of the world”\textsuperscript{61}, and so “good explanations...show how phenomena are situated within the causal structure of the world.” Thus, one contrast with the once-received view of scientific explanation- the Deductive-Nomological [D-N] model- is that whereas the D-N model focuses on discovering the laws of nature that can explain the phenomena at issue, the mechanistic model focuses on discovering the \textit{mechanisms} responsible.

The epistemic construal of mechanistic explanation is on board with the switch from laws to mechanisms, but differs from the ontic construal in that the epistemic construal emphasizes the different \textit{cognitive} activities that are required when representing and reasoning about mechanisms as opposed to laws. Thus, according to Bechtel, the D-N model requires that “the representations are propositional and the reasoning is logical inference”\textsuperscript{62}, whereas the epistemic construal emphasizes a different set of cognitive activities that are at play when representing and reasoning about the target phenomena in, for example, biology and psychology:

Representation and inference in mechanistic explanation is thus quite different from representation and inference in nomological explanation. While it is possible to give a linguistic description of a mechanism, the linguistic account is not privileged. Frequently diagrams provide a preferred representation of a mechanism. Inference involves a determination of how a mechanism behaves, and this is typically not achieved via logical inference but by simulating the activity of a mechanism, either by animating a

\textsuperscript{61} Craver (2007:21).

\textsuperscript{62} Bechtel (2008: 18).
diagram or by creating mental, computational, or scale model simulations. (Bechtel & Abrahamsen 2005: 432)\textsuperscript{63}

On the face of it, it isn’t clear why we should view Bechtel and Craver’s accounts as being in conflict with one another rather than as being two different, yet equally legitimate, facets of scientific explanation. For example, if we (i) provide a model describing the mechanism responsible for a given target phenomenon in neuroscience while (ii) emphasizing the cognitive activities involved in representing (not necessarily propositionally) and reasoning (not

\begin{footnotesize}
\textsuperscript{63} One issue that I don’t address in the main text is whether Bechtel’s approach to mechanistic explanation is really “epistemic” in the sense described by Salmon (1984; 1989). It isn’t clear what to say about this issue, because on the one hand Bechtel’s approach doesn’t seem to fall under either the inferential or erotetic versions of the epistemic conception (using Salmon’s terminology), yet on the other hand given the different approaches classified by Salmon as “epistemic” it isn’t entirely clear what, for Salmon, qualifies an approach to explanation as “epistemic”. Bechtel does emphasize epistemic concepts such as “representation” and “reasoning”, but these are not the concepts that were emphasized by earlier versions of the epistemic conception.

One way to interpret Bechtel’s approach as epistemic in Salmon’s sense is to follow Waskan’s (2006) discussion of Salmon’s opposition to the ‘psychological’ component of explanation and interpret the common element underlying Salmon’s epistemic conception as being the psychological component of explanation, aimed at clarifying “What is going on, cognitively speaking, when one has an explanation.” (2006: 201) So, while Bechtel’s approach doesn’t seem to fall under either the inferential or erotetic versions of the epistemic conception, if we interpret the epistemic conception as more broadly concerned with the psychological component of explanation, then it is correct to classify Bechtel’s approach as epistemic in the sense described by Salmon.
\end{footnotesize}
necessarily via logical inference) about the mechanism, it isn’t clear why/how (i) and (ii) are in conflict with one another.

However, Bechtel and Craver themselves explicitly present their accounts as being in conflict with one another. According to Bechtel, “taking ontic conceptions of mechanistic explanation at face value leads to absurdities”\(^\text{64}\), and he urges that “rejecting the D-N model...should not lead us to adopt an ontic account of explanation.”\(^\text{65}\) At the same time, Craver devotes most of the second chapter of (2007) towards championing the cause of the ontic construal by pointing out weaknesses in various strands of the epistemic conception of scientific explanation.

In the arguments for their respective positions, both Bechtel and Craver make use of intuitively correct claims about explanations: Bechtel identifies two intuitively correct claims about explanations that, he maintains, can be accommodated by the epistemic construal of mechanistic explanation but cannot be accommodated by the ontic construal of mechanistic explanation. Likewise, Craver has identified an intuitively correct claim about explanations that, he claims, can be accommodated by the ontic construal but cannot be accommodated by the epistemic construals. Here, then, are their arguments:

3.3.1. “There are Incorrect Explanations”

The first intuitively correct claim that, Bechtel argues, can be accommodated by the epistemic construal of mechanistic explanation but cannot be accommodated by the ontic

\(^{64}\) Wright & Bechtel (2007: 51).

\(^{65}\) Bechtel (2008: 18).
construal, is the claim that some explanations are incorrect. It is easy for Bechtel’s epistemic construal to accommodate the claim that there are indeed incorrect mechanistic explanations: The epistemic construal emphasizes the role played by the representations of mechanisms in mechanistic explanations, and representations can get things wrong and misrepresent the way the world actually is, i.e. be incorrect.

Bechtel argues that the ontic construal is unable to accommodate this claim because in cases involving incorrect mechanistic explanations, “…one has still appealed to a mechanism, but not one operative in nature.” The idea is that since the ontic construal identifies explanations with mechanisms, it is committed to the claim that the mechanism must actually exist, in order for an explanation to exist. However, if that is the case then it looks as though the ontic construal cannot take into account the existence of explanations that appeal to mechanisms that do not exist, i.e. incorrect explanations.

3.3.2. “Explanations are a Product of Epistemic Activity”

Bechtel argues that another way in which “taking ontic conceptions of mechanistic explanations at face value leads to absurdities” is that “characterizing explanation as non-epistemic is clearly problematic insofar as explanation is through-and-through an epistemic practice of making the world more intelligible by providing a rational basis for inferring how some given psychological phenomena is or could be produced.” Other formulations of this worry include: “Providing explanations, including mechanistic explanations, is essentially a


cognitive activity." And: “Explanation is fundamentally an epistemic activity performed by scientists.” Here the intuitively correct claim about explanations is that producing explanations requires some sort of cognitive or epistemic activity.

Once again it is easy for the epistemic construal to accommodate this claim, given its emphasis on representations and cognition. The epistemic features of mechanistic explanation that Bechtel emphasizes include “external representational devices” that are specific to mechanistic explanation, such as diagrams that can be used to represent the “multiple parts [of a mechanism] performing different operations, often in parallel with one another” and complex “causal or temporal relations in a mechanism” as well as types of reasoning “suitable to diagrammatic representations” that are specific to mechanistic explanation.”

When contrasted with the ontic construal, the critique is that the ontic construal cannot accommodate the claim that mechanistic explanation is fundamentally, essentially, or through-and-through an epistemic activity. This is because the ontic construal identifies explanations with mechanisms, and mechanisms per se are not epistemic activities.

3.3.3. “Explanations Exist Independently of Whether or Not They Are Discovered By Scientists”

Craver argues that the epistemic construal’s emphasis on the role played by cognitive abilities in explanations precludes the epistemic construal from being able to accommodate the intuitively correct claim that explanations exist independently of whether or not they are

discovered by scientists. As Craver writes, “The explanations exist even if we cannot represent them cognitively.”\textsuperscript{71} The epistemic construal appears to be committed to denying this intuitively correct claim when it pronounces that “explanation is fundamentally an epistemic activity performed by scientists.”\textsuperscript{72} If explanations are to be identified with epistemic activities then, ipso facto, their existence \textit{does} depend upon the discovery of certain facts by scientists.

3.3.4. \textit{What might the dispute be about?}

As I noted earlier, it isn’t entirely clear why we should think of the ontic and epistemic construals of mechanistic explanation as being in \textit{conflict} with one another, especially given the ease with which they can be presented as two different, yet equally legitimate, facets of scientific explanation. I think that a charitable interpretation of the dispute reveals two quite different possible disagreements: The first, which I shall refer to as the Uniqueness Dispute, presupposes that there is one and only one thing uniquely worthy of being called ‘explanation’, and the dispute then is over what the correct referent of ‘explanation’ actually is. The second, which I shall refer to as the Fundamentality Dispute, allows for more than one legitimate sense of explanation, but insists that one of the senses of explanation is more fundamental than the other sense(s). The dispute, then, is over which of the senses of explanation is the fundamental sense of explanation. In the next two sections I address, in turn, each of the two possible disagreements.

\textsuperscript{71} Craver (2007: 34).

\textsuperscript{72} Bechtel (2008: 18).
3.4. The Uniqueness Dispute

As we saw earlier, there is prima facie evidence in favor of there being multiple senses of the term ‘explanation’. In this section I reject the main assumption underlying the Uniqueness Dispute—namely, that there is one and only one thing uniquely worthy of being called ‘explanation’—by providing additional evidence in favor of ‘explanation’ being a polysemous term: First I reject several recent arguments made by Cory Wright in favor of ‘explanation’ only having one literal sense. Wright’s arguments turn on issues in lexical semantics, and it is worth pointing out here that it would surely be surprising if debates in the philosophy of explanation literature turned on issues in lexical semantics—after all, philosophers of science typically (should) see themselves as investigating the nature of explanation itself, rather than the term ‘explanation’. That said, for the sake of argument I grant the relevance of linguistics here and I argue that Wright gets the linguistics wrong. I then argue that both the ontic and epistemic senses are required in order to handle all three intuitively correct claims about explanations discussed in the previous section in the debate between Craver and Bechtel.

3.4.1. Rejecting Wright’s arguments in favor of ‘explanation’ only having one literal sense

In (2012) Wright presents several arguments in favor of ‘explanation’ only having one literal sense. In that paper Wright is mostly concerned with showing that ‘explanation’ is not ambiguous. However, everyone should agree that ‘explanation’ differs from paradigmatically ambiguous terms such as ‘bank’ insofar as the (putative) different senses of ‘explanation’ are clearly related to one another, whereas the different senses of ‘bank’ are clearly unrelated to
one another. On one common understanding of the relationship between ambiguity and polysemy\textsuperscript{73}, ambiguous terms whose senses are related to one another are referred to as ‘polysemous’, while ambiguous terms whose senses are \textit{unrelated} to one another are referred to as ‘homonymous’. Although unfortunately there are no widely agreed upon tests for polysemy\textsuperscript{74}, the \textit{distinction} between polysemous and homonymous terms is (relatively) widely accepted. Furthermore, and importantly for our purposes, tests for homonymy are generally thought to \textit{not} work as tests for polysemy. Thus, paradigmatically polysemous terms such as ‘bottle’ (which can refer both to a container and to a quantity) typically \textit{fail} tests for homonymy\textsuperscript{75}.

Now, Wright claims that ‘explanation’ is not in fact lexically ambiguous, and he presents linguistic evidence in arguing for this claim by pointing out that although ‘explanation’ passes one of the tests for lexical ambiguity, it fails some of the other tests. However, we must be wary of tests for lexical ambiguity that fail to distinguish polysemy from homonymy. As I shall show, some of the tests for lexical ambiguity are really tests specifically for \textit{homonymy}, and it is thus perfectly understandable why terms like ‘explanation’ could fail these tests yet still be ambiguous, just not homonymous. Unfortunately, after acknowledging that ‘explanation’ passes one of the tests for lexical ambiguity, Wright repeatedly makes the mistake of inferring that ‘explanation’ is not ambiguous on the basis of its failing to pass a test for homonymy.

3.4.1.1. The syllepsis/conjunction reduction test

\textsuperscript{73} Taken from Cruse (2004).

\textsuperscript{74} For further discussion see \textit{inter alia} Geeraerts (1993); Petho (2001); Cruse (2004).

\textsuperscript{75} Ibid.
First, though, it is important to point out that the test that ‘explanation’ does pass - the syllepsis/conjunction reduction test - is better suited for detecting polysemy rather than homonymy. Consider the following unacceptable (and thereby passing the syllepsis/conjunction reduction test) sentence, taken from Wright (the unacceptability is indicated with an asterisk):

(5*) The explanation of why the ValuJet crash in the Florida Everglades occurred was given by the investigator and the oxygen canisters.

This seems to be a perfect example of referencing multiple related senses of a term with a single token use of the term. The following two sentences, of course, are acceptable:

(5a) The explanation of why the ValuJet crash in the Florida Everglades occurred was given by the investigator.

(5b) The explanation of why the ValuJet crash in the Florida Everglades occurred was given by the oxygen cannisters.

What is unacceptable is to put (5a) and (5b) together in the same sentence in the manner above (5*). Notice, though, that in order for both (5a) and (5b) to be separately acceptable, it must be the case that the different senses of ‘explanation’ are closely related. For if the ontic and epistemic senses of ‘explanation’ were unrelated, just like the different senses of ‘bank’, then it would be difficult to construct a sentence such as (5*) in the first place, where two acceptable sentences are being conjoined. Thus, the fact that ‘explanation’ passes the
syllepsis/conjunction reduction test constitutes stronger evidence in favor of ‘explanation’ being polysemous than Wright gives it credit for.

3.4.1.2. The Test of Contradiction

Paradigmatically ambiguous (or better: homonymous) terms such as ‘bank’ pass the test of contradiction insofar as they can be featured in perfectly acceptable sentences such as the following:

(6a) That bank is not a bank.

However, the same type of sentence becomes unacceptable when ‘bank’ is replaced with ‘explanation’:

(6b) That explanation is not an explanation.\(^76\)

However, on the hypothesis that ‘explanation’ is polysemous rather than homonymous, we should not be surprised that unlike ‘bank’, which is a homonymous term, ‘explanation’ fails

\(^76\) Wright himself uses the sentence “No non-explanation is an explanation”, but I think that the sentence “That explanation is not an explanation” is both more closely analogous with “That bank is not a bank” and is more straightforwardly unacceptable than “No non-explanation is an explanation”, and hence on both counts better suited for Wright’s purposes. He also, confusingly, uses the sentence “To explain is not to explain” as an additional disanalogy to “That bank is not a bank.” However, the relevant ambiguity at issue is between the representation (i.e. the epistemic referent) and its content (i.e. the ontic referent), rather than between the act of explaining and a state of affairs. So the (un)acceptability of “To explain is not to explain” is irrelevant for our purposes.
the test of contradiction. Since there is no relationship amongst the different senses of ‘bank’, it is easy to understand why a sentence such as (6a) is acceptable. But since there is, ex hypothesi, a close relationship amongst the different senses of ‘explanation’, it is similarly easy to understand why a sentence such as (6b) is unacceptable. Failing the test of contradiction is evidence against ‘explanation’ being homonymous, but it does not count as evidence against ‘explanation’ being polysemous. It is simply a mistake to infer, as Wright does, that ‘explanation’ is not polysemous on the basis of its failing to pass a test for homonymy.

3.4.1.3. Test checking for two or more unrelated antonyms

A paradigmatically ambiguous (or better: homonymous) terms such as ‘light’ is ambiguous between ‘not heavy’ and ‘not dark’. Wright argues that since ‘explanation’ does not similarly possess multiple unrelated antonyms, it fails the test that checks for two or more unrelated antonyms. Again, though, failing this test may count as evidence in favor of a term being ambiguous with multiple unrelated senses- i.e. homonymous-, but it would make no sense to say that failing this test counts as evidence in favor of a term being ambiguous with multiple related senses- i.e. polysemous. After all, if a term has multiple related senses, why should we expect the term to have two or more unrelated antonyms? Once again, it is simply a mistake to infer, as Wright does, that ‘explanation’ is not polysemous on the basis of its failing to pass a test for homonymy.

3.4.1.4. Rebutting further claims made by Wright

In addition to the tests for linguistic ambiguity, Wright also claims that (a) the ontic sense of explanation is a merely metaphorical sense, and (b) ‘explanation’ cannot be
polysemous, as it is quite unlike other polysemous terms. However, Wright doesn’t do much by way of argument for these claims. Starting with (a): According to Wright, the ontic sense of explanation is akin to metaphorical statements such as

(7) My car refuses to start

Thus, “mechanisms explain in precisely the same sense that cars are moody or have poor etiquette, that [dopamine] systems beg, borrow, and steal, or that natural selection is a shrewd negotiator.” (Wright 2012: 388)

The problem, though, is that Wright gives no argument in favor of this metaphorical interpretation of the ontic sense of ‘explanation’, and there is no reason to think that we should interpret the ontic sense of ‘explanation’ as metaphorically attributing intentional behavior (i.e. intentionally providing explanations) to mechanisms.

Onto (b): There is a short, and puzzling, passage in (2012) where Wright addresses the possibility of ‘explanation’ being a polysemous term. Here is the passage that contains the entirety of Wright’s discussion of this issue:

...is it not just an empirical fact that explanation is polysemous, having multiple but closely related senses? No... It is certainly not polysemous in the way that some terms are; e.g., prepositions such as of, around, and over are well-known to have a wide array of senses, even as many as one hundred or more. Nouns (e.g., ring, which includes the circular mark, circular piece of jewelry, group of people operating together clandestinely, boxing arena, etc.) and verbs (e.g., nurse, which includes senses such as breast-feed, care for the infirm, sip infrequently) often have far fewer senses. (Wright 2012: 385-6)
We can of course agree with Wright that ‘explanation’ has far less senses than do polysemous prepositions. We can also agree with Wright that there are other polysemous terms that have far less sense senses than do polysemous prepositions. The strongest conclusion warranted by this passage is that ‘explanation’ is not a polysemous preposition. I agree, but of course no one thinks that ‘explanation’ is a preposition. Absent an (absurd) argument that polysemous terms must have at least one hundred different senses, Wright has not provided any evidence here refuting the claim that ‘explanation’ is a polysemous term (albeit not, of course, a polysemous preposition).

3.4.2. The intuitively correct claims revisited

On the Uniqueness Dispute interpretation of the EVO debate, the intuitively correct claims about explanations discussed above can be used as follows: The intuitively correct claims about explanations that (Bechtel maintains) can be accommodated by the epistemic construal of mechanistic explanation but cannot be accommodated by the ontic construal should be seen as evidence in favor of preferring the chosen referent of the E-theorists. Likewise, the intuitively correct claim about explanation that (Craver maintains) can be accommodated by the ontic construal but cannot be accommodated by the epistemic construal should be seen as evidence in favor of preferring the chosen referent of the O-theorists.

In this section I argue that the only way to accommodate all three intuitively correct claims about explanations is by acknowledging that ‘explanation’ is a polysemous term. In other words, both the ontic and epistemic senses are required if we wish to accommodate all three intuitively correct claims. Thus, rather than providing an argument in support of the O-theorists
and in opposition to the E-theorists, or vice versa, the intuitively correct claims about explanations provide another source of evidence in favor of ‘explanation’ being a polysemous term.

3.4.2.1. “There are Incorrect Explanations” Revisited

Recall that the first intuitively correct claim about explanations that, Bechtel maintains, can be accommodated by the epistemic construal but not by the ontic construal, is the claim that some explanations are incorrect. It would be very difficult to try to make sense of scientific practice without acknowledging that some explanations are incorrect. After all, the whole notion of scientific progress is predicated upon the assumption that science aims to build upon and correct previous explanations that have been offered. Bechtel is correct insofar as accommodating this intuitively correct claim about explanations requires acknowledging the existence of an epistemic sense of explanation— that is, a sense of explanation that refers to the representations individuated by the content of explanatory products. Acknowledging the existence of an epistemic sense of explanation should not, of course, imply that the epistemic sense of explanation is the only sense of explanation. Thus, New Mechanists who acknowledge the existence of both the epistemic and ontic senses of explanation should be willing to endorse the following claims: Sometimes people explain things correctly, and sometimes people explain things incorrectly. And when people explain things incorrectly, they may be appealing to nonexistent mechanisms. Likewise, sometimes the representations individuated by the content of explanatory products— such as the content of explanatory texts— are correct, and sometimes they are incorrect. And when the representations are incorrect, they are either mistakenly referring
to non-existent mechanisms, or mistaken about the relationship that actually obtains and results in the phenomenon in question.

3.4.2.2. “Explanations are a Product of Epistemic Activity” Revisited

Recall that the second intuitively correct claim about explanations that, Bechtel maintains, can be accommodated by the epistemic construal but cannot be accommodated by the ontic construal, is the claim that producing explanations requires some sort of cognitive or epistemic activity. This is a fairly obvious claim concerning all explanations, scientific or otherwise, and again Bechtel is correct insofar as accommodating this claim requires acknowledging the existence of an epistemic sense of explanation. Again, though, acknowledging the existence of an epistemic sense of ‘explanation’ should not imply that the epistemic sense of explanation is the only sense of explanation. Thus, New Mechanists who acknowledge the existence of both the epistemic and ontic senses of explanation can happily grant that there is epistemic activity involved in mechanistic explanations- that is, that people do give mechanistic explanations (these are the explaining acts), and that there are products of these explaining acts that possess representational content. Likewise, there is no reason to think that New Mechanists who acknowledge the existence of both the epistemic and ontic senses of explanation are somehow incapable of taking into account the particular vehicles of content that Bechtel emphasizes, such as diagrams and visual images.

3.4.2.3. “Explanations Exist Independently of Whether or Not They Are Discovered By Scientists” Revisited

80
Recall that the intuitively correct claim about explanations that, Craver maintains, can be accommodated by the ontic construal but cannot be accommodated by the epistemic construal, is the claim that explanations exist independently of whether or not they are discovered by scientists. This claim is important for any non-anthropocentric understanding of scientific practice, whereby an important goal of scientific practice is to discover the explanations for phenomena, and the existence of the explanations does not require or presuppose the existence of, let alone discovery by, scientists or any other cognitive agents. Craver is correct insofar as accommodating this intuitively correct claim about explanations requires acknowledging the existence of an ontic sense of explanation— that is, a sense of explanation that refers to explanatory states of affairs. Acknowledging the existence of an ontic sense of explanation should not, of course, imply that the ontic sense of explanation is the only sense of explanation. Thus, New Mechanists who acknowledge the existence of both the ontic and epistemic senses of ‘explanation’ needn’t deny the existence of explanatory states of affairs that could be picked out by representations. Thinking back to our earlier example, then, New Mechanists who acknowledge the existence of both the ontic and epistemic senses of explanation needn’t deny the fact that in the 18th century, with respect to action potentials, there existed referents, and a causal relationship, that could be picked out by a representation individuated by the content of an explanatory product.

Here, then, is the upshot of our discussion concerning the intuitively correct claims about explanations: O-theorists who deny the existence of an epistemic sense of explanation cannot accommodate the claim that there are incorrect explanations, and they cannot accommodate the claim that explanations are a product of epistemic activity. E-theorists who
deny the existence of an ontic sense of explanation cannot accommodate the claim that explanations exist independently of whether or not they are discovered by scientists. Thus, both the epistemic and ontic senses of explanation are required in order to accommodate all three intuitively correct claims about explanations. And since the rejection of any of these intuitively correct claims about explanations would fly in the face of trying to make sense of scientific practice, it follows that the intuitively correct claims about explanations provide additional evidence in support of ‘explanation’ being a polysemous term with (at least) both an epistemic and an ontic sense.

3.5. The Fundamentality Dispute

The debate between Craver and Bechtel can also be interpreted as a dispute over which of the senses of explanation is the fundamental sense of explanation. Thus, according to Phyllis Illari’s recent assessment of the debate,

While some arguments are still offered in the style of the classic debate—linguistic analysis of explanation (Wright 2012), or using cases of paradigm explanations to assess ascription conditions for ‘explanation’ (Wright 2012; Craver 2012)—a new line of argument is coming to prominence... in this [new] frame Craver is not arguing about what an explanation itself is, but arguing for the importance of ontic constraints in recognizing, finding, and possibly even using good explanations... Bechtel, and Bechtel and Wright, are not so explicit, but they recognize and are concerned about such norms: ‘explaining refers to a ratiocinative practice governed by certain norms’ (Wright and Bechtel 2007: 51), although they also sometimes appear to be arguing about what mechanistic explanations themselves are. It is true that neither side consistently sticks to arguing over constraints, sometimes changing tack to argue in a more traditional way over what
explanations themselves are. However, they are moving towards a focus on constraints, away from the traditional ontic-epistemic debate. (Ilari 2013: 240-1)

On this interpretation, then, O-theorists claim that the ontic sense of explanation is fundamental in virtue of certain norms of scientific explanation following from the ontic sense, whereas E-theorists claim that the epistemic sense of explanation is the fundamental sense of explanation. In this section I argue that Craver’s argument for fundamentality is unsuccessful—indeed, there is reason to think that any such argument on behalf of O-theorists or E-theorists—will be similarly unsuccessful.

Craver (2013) argues that “an appeal to ontic explanation is essential for marking several crucial normative dimensions by which scientific explanations are and ought to be evaluated”:

- The distinction between how-possibly and how-actually-enough explanations
- The distinction between phenomenal descriptions and explanations
- The difference between predictive and explanatory models
- The requirement that explanatory models should include all and only information that is explanatorily relevant to the phenomenon one seeks to explain

Thus, good scientific explanations tell us not just how something might have happened, but how it actually happened; they tell us not just what happened but why it happened; they don’t just make predictions, for predictive success can occur in the absence of getting things right about the world; and they tell us exactly, and only, what it is in the world that is “explanatorily relevant to the phenomenon one seeks to explain.”

The first problem with Craver’s fundamentality argument is that, as Ilari (2013) notes, Craver has (at most) shown that the ontic sense of ‘explanation’ is important insofar as it
provides us various important constraints on good scientific explanations. But this doesn’t rule out the possibility that the epistemic sense of ‘explanation’ is also important insofar as it provides us with other important constraints on good scientific explanations—such as, perhaps, simplicity, coherence, unification, etc. Thus, even if Craver has shown that the ontic sense of ‘explanation’ is essential for determining some of the norms of scientific explanation, he has failed to show that the ontic sense is fundamental insofar as he has failed to show that the epistemic sense is not essential for determining any of the norms of scientific explanation.

The second problem with Craver’s fundamentality argument—and, for that matter, any fundamentality argument revolving around the norms of scientific explanation—is that, on closer inspection, none of the norms of scientific explanation “follow” from one sense of explanation without equally following from the other sense. Thus:

(a) It is easy to translate the norms of scientific explanation that “follow” from the ontic sense as equally following from the epistemic sense; and

(b) It is easy to translate the norms of scientific explanation that “follow” from the epistemic sense as equally following from the ontic sense.

Starting with (a): The epistemic sense of explanation typically emphasizes the importance of good scientific explanations not only getting things right about the world, but also increasing our understanding of the world. Thus, an E-theorist can similarly make the following claims about requirements on good scientific explanations: Good scientific explanations increase our understanding by

- Telling us not just how something might have happened, but how it actually happened
- Telling us not just what happened but why it happened

84
• Not just making predictions, for predictive success can occur in the absence of any increase in genuine understanding

• Including “all and only information that is explanatorily relevant to the phenomenon one seeks to explain”

Continuing with (b): No E-theorist has (as of yet) argued for the counterpart of Craver’s (2013) claim. That is to say, no E-theorist has argued that “an appeal to epistemic explanations is essential for marking several crucial normative dimensions by which scientific explanations are and ought to be evaluated.” Thus, there is no extant list of norms of scientific explanation that are alleged to follow from the epistemic sense, and hence no list of norms to translate as equally following from the ontic sense. Nonetheless, I hypothesize that any such list of norms will indeed be easy to translate as equally following from the ontic sense, given that E-theorists typically emphasize the importance of good scientific explanations increasing our understanding of the world. Thus, any norm of scientific explanation that is alleged to follow from the epistemic sense should, in theory, be translatable as equally following from the ontic sense, on the (ontic) grounds that good explanations that increase our understanding of the world are only good insofar as they get things right about the world. Thus, for example, the use of diagrams may (in certain cases) be required in order to get things right about the world. In which case, the claims of E-theorists (such as Bechtel) who emphasize the important role of diagrams in mechanistic explanations can easily be translated as following from the ontic sense: Vehicles of representation (such as, in some cases, diagrams) that get things right/more right about the world are to be preferred to vehicles of representation (such as, in some cases, propositions) that get things wrong/less right about the world.
Further reflection on Craver’s argument here concerning the relationship between explanatory norms and fundamentality reveals an underlying basic confusion. Let’s say that I have the following thought: ‘The cat is on the mat’. In order for that thought to be a true thought, it must actually be the case that there is indeed a cat on the mat. Does this show that thoughts are “fundamentally” states of affairs as opposed to, say, mental particulars? The answer, I take it, is obviously no. Yet it is precisely this line of reasoning that Craver is attempting to use in arguing from the premise that various explanatory norms follow from the ontic sense of explanation to the conclusion that the ontic sense of explanation is the fundamental sense of explanation. If the fact that the truth of ‘the cat is on the mat’ depends on worldly states of affairs can’t get us to the conclusion that thoughts are fundamentally states of affairs, then so too the fact that the truth of scientific explanations (i.e. at least part of their goodness) depends on worldly states of affairs can’t get us the conclusion that explanations are fundamentally ontic.

3.6. Diagnosing the debate

We have seen that the term ‘explanation’ has several closely related senses, and is not ambiguous in the same way that terms such as ‘bank’, which have several unrelated senses, are. We have also seen that it is quite easy to translate claims that seem to follow from one sense of explanation into claims following from a different sense of explanation. Finally, we have seen that no good arguments have been offered in support of the claim that the different senses of explanation are in conflict with one another. In this final section I provide a simple explanation
for the above *explananda*, which at the same time shows how the EVO debate is deeply confused insofar as it contains a pseudoproblem.

The simple explanation for the above *explananda* is that ‘explanation’ is a polysemous term with multiple (though related) legitimate senses, and the relationship between the different senses of explanation mirrors the relationship between thoughts and the world. It is widely appreciated that in many circumstances thoughts are individuated by their referential contents, which in turn needn’t require cognizers at all. In other words, the individuation conditions on thoughts have the following interesting property: Thoughts (and, more generally, representations) can be individuated by things—namely, their referential contents—that don’t require the existence of that which is being individuated—namely, the thoughts themselves or the representations themselves. Similarly, explanations *qua* epistemic sense can be individuated by explanations *qua* ontic sense, and explanations *qua* ontic sense don’t require the existence of explanations *qua* epistemic sense.

Thus, once we see that the relationship between the different senses of explanation mirrors the relationship between thoughts and the world, it is easy to see that the EVO debate is deeply confused. The EVO debate presupposes that the different senses of explanation are in conflict with one another. But this makes as much sense as worrying that thoughts about the world must either be seen primarily as *thoughts* about the world, or instead seen primarily as thoughts about *the world*.

I conclude that the most recent version of the EVO debate should be resolved in much the same way that Salmon resolved the original version of the debate in 1989: The ontic and epistemic construals of mechanistic explanation should be seen as reflecting different (though
related) legitimate senses of explanation, and there is no reason to think that the different senses of explanation are in conflict with one another.
Chapter 4: Reasons, Causes, and the Extended Mind Hypothesis (with Timothy Schroeder)

The extended mind thesis (EMH) has generated much controversy ever since the publication of Andy Clark and David Chalmers’ “The Extended Mind” in 1998. While its critics have been numerous and vocal, the debate appears to have reached a stalemate, insofar as the criticisms rely upon premises not shared by proponents of EMH. One recent suggestion, then, has been to use the New Mechanists’ mutual manipulability (MM) account of constitutive relevance in order to allow for a non-question begging, and empirically based, means of resolving the debate over extended cognition in favor of EMH.

In this paper we develop a novel argument against EMH, employing only premises that proponents of EMH, as well as the New Mechanists, have independent reasons to accept. Our argument makes use of a necessary condition on holding a belief for a reason, and then shows that many of the external objects that defenders of EMH have claimed are (in conjunction with an internal cognitive system) beliefs cannot (even partly) constitute beliefs had for a reason. Our first thesis, then, is that beliefs had for reasons cannot be extended in the most interesting

77 See, inter alia, Rupert (2004; 2009); Adams & Aizawa (2008); Weiskopf (2008); and Spaulding (2012).

78 See Coleman (2011) for another discussion of the idea that arguments made by different parties in the debate have involved begging the question against the other side.

79 This necessary condition was developed by Arpaly (2006) and Wedgwood (2006), independently of one another. For a slightly different version of the necessary condition see Turri (2011).

80 Our argument is not directed against proponents of EMH who focus on socially distributed cognition, rather than external objects and artifacts. For further discussion see Sutton et al. (2010).
way specified by EMH. Furthermore, if we add to our argument the premise that every token
belief is something that could have been had for a reason, we can derive the stronger thesis that
no beliefs can be extended in this way. Thus, while the MM account of constitutive relevance
may be useful for demarcating the boundaries of mechanisms studied in many of the sciences, it
cannot be used to support EMH.

In 4.1 we present EMH, in 4.2 we rehearse the debate between its critics and defenders,
and in 4.3 we present the proposal on behalf of EMH that makes use of the MM account of
constitutive relevance. In 4.4 we introduce our necessary condition on having a belief for a
reason, in 4.5 we use this necessary condition to provide our novel argument against EMH, and
in 4.6 we respond to a possible objection.

4.1. The Extended Mind Hypothesis

According to EMH, the physical vehicles of at least some of a person’s mental states can
be located outside of the person’s body. That is, external objects can (at least partly) constitute
at least some mental states, or at least some cognitive processes. Thus, according to Clark and
Chalmers’ canonical statement of the view, “beliefs can be constituted partly by features of the
environment, when those features play the right sort of role in driving cognitive processes.”81

81 Clark and Chalmers (1998: 12). For more recent book-length articulations of EMH, see Clark (2008);
Menary (2007); and Rowlands (2010).
The extended mind hypothesis can also be called ‘transcranialism’, as opposed to the ‘intracranial’ position according to which the mind is found wholly within the skull.\textsuperscript{82}

After more than a dozen years of debate, it seems clear that the real dispute is not over transcranialism as such, because there are certain scenarios in which even the most vocal critics of Clark and Chalmers would agree that part of a human being’s mind is located outside of her skull. For a simple example, just imagine that some region of a person’s brain is removed from her skull and kept alive in a jar she carries around with her.\textsuperscript{83} And imagine that the functionality of all the disrupted neural connections is maintained by tiny radio transceivers and attached effectors. These tiny devices ensure that, when a given neuron would have received a chemical signal when the brain was intact, the chemical receptors of that neuron act just as though they have received exactly that chemical signal. Thus, the functional architecture of the person’s brain is preserved down to almost the smallest scale possible. The part of the brain inside the skull interacts seamlessly with the part of the brain outside the skull. There are a few lingering functional changes (consuming alcohol only affects part of the person’s brain, now) but these functional changes do not strike participants in the debate as the sort of thing that is genuinely relevant to mentality. Everyone in the debate is ready to agree that part of this person’s mind – the vehicles for some of her thoughts, desires, and the like – are found outside of her skull.

\textsuperscript{82} According to some self-described ‘embodied cognition’ theories the mind is not restricted to the central nervous system, but does stop at least at the skin of the body. This intermediate position is one we will neglect in this paper, since it will not be relevant to our arguments. Thus, using the helpful taxonomy of Sutton et al. (2010: 531), we are addressing extended cognition, rather than embedded cognition or distributed cognition.

\textsuperscript{83} The obvious antecedents of our thought experiment are found in Dennett (1978) and Nagel (1971).
The real dispute is over an ill-defined class of cases centered on a paradigmatic and controversial case introduced by Clark and Chalmers: the case of Inga and Otto. Inga has a belief concerning the location of New York’s Museum of Modern Art – it is on 53rd Street – that she can consult by searching her memory. Otto suffers from Alzheimer’s disease and so does not have the luxury of consulting his memory in the ordinary way. Instead, he consults a notebook he carries around with him: a notebook that allows him to replicate certain functional features of ordinary human memory. For instance, when Otto learns the location of New York’s Museum of Modern Art, he writes this location down in his notebook (just as Inga would add this fact to her memory), and when he wonders about the location of MoMA he consults his notebook (just as Inga would consult her memory). Clark and Chalmers infamously argue that it is possible that Otto’s case is one in which an inscription in Otto’s notebook, such as the one noting the location of MoMA, has sufficient functional similarities to ordinary memories (such as Inga’s) about the location of MoMA, and so it is possible for Otto to be a person who satisfies EMH. So long as Otto automatically, habitually consults his notebook and automatically, habitually records information in it, so long as Otto unthinkingly trusts the writings in his notebook (to the extent that healthy people unthinkingly trust their own memories), and so long as Otto retains his eyesight, capacity to write, capacity to read, and so on – in short, so long as Otto implements a good functional approximation to the gross functional architecture of memory storage and use – Otto will count as having some of his beliefs, and so some of his mind, in his notebook. The inscriptions in the notebook will be among the vehicles of his beliefs.

4.2. Some criticisms of EMH as applied to Otto

Critics have attacked the thesis that Otto’s notebook inscriptions are among his belief tokens – that the notebook is a part of Otto’s mind – from either of two main directions. But neither direction of attack has been particularly successful.

One kind of attack has focused on foundational issues in the philosophy of mind. For instance, Fred Adams and Ken Aizawa have argued that the inscriptions in Otto’s notebook have only derived intentionality, while beliefs have original intentionality.\(^ {85}\) Representations with derived intentionality, such as “words, stop signs, warning lights and gas gauges mean what they do through some sort of social convention,”\(^ {86}\) whereas beliefs and other mental representations have original intentionality in that they “mean what they do in virtue of satisfying some naturalistic conditions on meaning.”\(^ {87}\)

The problem with this particular attack is that it is far from obvious that the inscriptions in Otto’s notebook have only derived intentionality. As Clark argues:

We are not required to think of Otto’s notebook as constraining some plausible story about intrinsic content...just because the symbols in the notebook happen to look like English words and require some degree of interpretative activity when retrieved and used, that need not rule out the possibility that they have also come to satisfy the demands on being, in virtue of their role within the larger system, among

\(^ {85}\) See Dretske (1988, chapter 3) for one particularly clear presentation of this distinction.

\(^ {86}\) Adams and Aizawa (2010: 70).

\(^ {87}\) Ibid.
the physical vehicles of various forms of intrinsic content [i.e., original intentionality]...I do not believe that there is any nonquestion-begging notion of intrinsic content that picks out all and only the neural in any clear and useful fashion. (Clark 2008: 90)

Functionalism of the sort Clark and Chalmers endorse permits the inscriptions in Otto’s notebook, and indeed any physical object whatsoever, to count as having original intentionality so long as they play the right functional roles. To simply deny that the inscriptions have this kind of intentionality without making the case that the inscriptions play the wrong functional role is thus just to reject the kind of functionalism that Clark and Chalmers embrace, and this is not a rejection that will move defenders of EMH.

The problem with this specific objection is one that threatens any similarly foundational objection to the claim that Otto’s mind is found partly in his notebook. Any foundational objection is in danger of taking as a premise some thesis about the mind that will be rejected by any functionalist of the Clark and Chalmers sort. And any such objection will then either beg the question, or turn the debate into one about the truth of the sort of functionalism that Clark and Chalmers endorse. And either way, nothing has been gained by thinking specifically of the case of Otto.

The other kind of attack has focused on the specific functional details of Otto’s notebook use, pointing out that there are many functional differences between it and the operation of memory in ordinary healthy human beings. These attacks accept the basic functional framework of Clark and Chalmers, but argue that the functional details of Otto’s notebook use are the wrong functional details to reach the striking EMH result in Otto’s case. One of the more conservative instances of this strategy has been exemplified by Adams and Aizawa. They point out that there are certain laws of normal human biological memory
processes, such as the primacy effect and negative transfer, that are not laws governing Otto’s use of his notebook. Thus, there are functional disanalogies between ordinary healthy memory and Otto’s use of his notebook, and so there are grounds for rejecting the claim that Otto’s memories are found in his notebook.

A less conservative, more general instance of the strategy comes from Daniel Weiskopf. He argues that unless beliefs are part of a system whereby information is integrated via belief updating mechanisms, they aren’t really beliefs. For example, if Inga has a belief that MoMA is on 53rd Street, and later comes to learn that MoMA is being torn down, her beliefs will be updated such that she no longer believes that MoMA is on 53rd Street – perhaps she will instead believe that MoMA was on 53rd Street. Moreover, this type of information integration is ubiquitous and occurs (at least typically) rapidly and effortlessly. In contrast, if Otto writes down in his notebook that MoMA is on 53rd Street, and then later writes down that MoMA is being torn down, the inscriptions in the notebook do not undergo a similar process of information integration. The notebook inscription that MoMA is on 53rd Street stays the same, even though there is another inscription in the notebook that says that MoMA is being torn down. This pervasive functional disanalogy is said to block the claim that Otto’s notebook contains some of his beliefs.

The most general instance of the strategy argues that the successes of current cognitive science should dictate which states and processes are properly labeled ‘cognitive’ and which are not. Thus, Robert Rupert argues that the successes of current cognitive science are best


89 Weiskopf (2008).
explained by the actual existence of a construct presupposed by cognitive scientists, namely, “a persisting, relatively unified, and organismically bounded cognitive system.”

Cognitive phenomena are produced distinctively by the persisting collections of capacities, mechanisms, skills, and abilities that play a privileged role in successful cognitive psychological explanation...A state is cognitive iff it consists in, or is realized by, the activation of one or more mechanisms that are elements of the integrated set members of which contribute causally and distinctively to the production of cognitive phenomena. (Rupert 2009: 38, 42)

The functional role played by Otto’s notebook is implemented via intermediate capacities, mechanisms, skills, and abilities (the ability to see the notebook, the ability to move his hand to write, habits of writing, habits of consulting the notebook, and so on) that are not those deployed by ordinary people consulting their memories. Thus, his notebook use is not an instance of one or more of the “capacities, mechanisms, skills, and abilities” that psychologists appeal to in their current psychological explanations of memory. As a result, Rupert holds, the states of Otto’s notebook cannot be memories.

One way to understand the debate between Clark and Chalmers and the above critics is to take the former to be embracing a coarse-grained commonsense functionalism and the latter to be embracing a fine-grained psychofunctionalism. Understood in this way, these criticisms again fail to be specific to the details of EMH, and instead go straight to deeper issues about the nature of the mind.

So suppose, instead, that the above critics can be thought to share with Clark and Chalmers a general commitment to functionalism but not a specific commitment to a particular sub-type of functionalism. The problem with all three of these functional-difference arguments

---

against the EMH interpretation of Otto is that they all make a great deal of the functional differences they identify without having a principled basis for doing so. The defender of Otto as an exemplification of EMH will insist that, in spite of the admitted functional differences, there are nonetheless more important functional similarities between Otto and the ordinary healthy person: functional similarities at a more coarse-grained level. Clark and Chalmers and their allies have suggested a “Parity Principle” that can be understood91 as defending their more coarse-grained functionalism. According to Clark’s (2010) version92 of the Parity Principle:

If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process.

While it is true that the fine-grained functional details of Otto’s notebook use differ from the fine-grained functional details of ordinary, healthy memory use, Clark and Chalmers contend that if we would encounter someone (perhaps an alien) with a very Otto-like information processing system located entirely within his, her, or its skull (carapace, membrane...) then we would grant that we were encountering a memory system of a distinctive kind. We would admire the ingenuity of nature, perhaps: it gives rise to organisms with broadly the same functional capacities (to store and recall apparent facts) while finding distinctive ways for these different organisms to implement this capacity. It would be anti-functionalist speciesism to deny an alien with an internal but otherwise Otto-like information processing

91 This way of thinking about the dialectic between critics and defenders of EMH is taken from Spaulding (2012), and is perhaps a bit idiosyncratic, especially when it comes to describing the role played by the Parity Principle. However, nothing in our argument hinges upon this way of thinking about the dialectic.

92 Clark (2010: 44).
system the claim to having memories, it seems. And thus, Clark and Chalmers can conclude, it is mere intracranial prejudice to insist that Otto, who is just like the alien except in the location of his putative memories, does not have the memories that his gross functional behavior suggests. In response, the philosophers making functional-role attacks on the Clark and Chalmers interpretation of Otto might try to hold that the functional differences to which they personally have pointed are functional differences that really matter, but such responses do not strike us as wholly convincing. The narrowest attack pointed out that Otto’s notebook is not subject to phenomena such as primacy effects or negative transfer, but it is hard to believe that if we were to meet a human being (or an alien) who was not susceptible to these effects, but who was otherwise normal, that we would deny that the individual in question had a memory; a generic sort of functionalism suggests a more generous interpretation.

The next attack pointed out Otto’s problems with “information integration.” But, as Spaulding (2012: 16) points out:

Otto’s beliefs and notebook entries are part of a system with belief updating mechanisms. The updating mechanisms are precisely the same mechanisms that update beliefs in the normal case. It is just that in Otto’s case the mechanism operates more slowly and is implemented differently sometimes. We can add that, in human beings with real memory-impairing conditions such as Alzheimer’s disease, there are also widespread failures in belief updating. There is often some purely neural belief updating, of the sort that Otto can be imagined to implement moderately well via his notebook. But the purely neural updating is far from perfect, which would also be true of Otto with his notebook. Since real human beings with Alzheimer’s disease have biological memories (or at least seem to, from a commonsensical perspective), though very imperfect ones, it seems
difficult to hold that Otto’s imperfections in belief-updating (relative to healthy human beings) entail just in themselves that the marks in his notebook are not memories.

The most general attack pointed out that Otto’s capacities, mechanisms, skills, and abilities involved in using his notebook are distinct from those appealed to in actual scientific explanations of memory storage and retrieval. However, this is at least partly due to the fine-grained details in functional implementation. If psychologists would encounter aliens with Otto-like but internal information storage and retrieval processes, they might add to the range of psychological explanations they deployed when discussing memory. It would not do for a critic of Clark and Chalmers to merely assume that the laws and mechanisms of a future, fully developed cognitive psychology will not appeal in its explanations of memory to the capacities and abilities of such aliens. And if it cannot be assumed that these capacities and abilities, found intracranially, will never be part of cognitive psychology then it cannot be assume that these same capacities and abilities, found transcranially (in Otto’s case), will never be a part of cognitive psychology.  

In the light of these unsuccessful attacks on the Clark and Chalmers interpretation of Otto’s case, it seems that one can draw a tentative moral: a more promising strategy for refuting EMH might be one that both starts from premises that are acceptable to a coarse-grained functionalist and that does not rely on functional disanalogies between Otto and ordinary human beings that are just a matter of fine detail or degree. First, though, in the next section we present one strategy that has been offered on behalf of EMH that begs no questions against EMH’s opponents.

---

93 This criticism is not original to us. Amongst others, Spaulding (2012) makes a similar point.
4.3. A mechanistic solution to the extended cognition debate

David Michael Kaplan (2012) notes that the issue of disagreement between proponents of EMH and its critics can be seen as concerning the proper demarcation criteria for cognition—the criteria that determines whether something should, or should not, count as part of a cognitive process. Kaplan also points out that the criteria that have been offered have all been criteria proprietary to cognition—that is, rather than being criteria for process, or system, membership in general, the criteria are specific to cognitive processes or systems.

Kaplan’s proposal, then, is to borrow the independently motivated New Mechanist account of constitutive relevance—which specifies criteria for mechanism, or system, membership in general—and apply these criteria to specifically cognitive mechanisms or systems. According to the MM account, if X is a component of Y, then (roughly) if we wiggle X we will see a corresponding change in Y, and if we wiggle Y, we will see a corresponding change in X. A bit more formally:

X’s $\Phi$-ing is constitutively relevant to S’s $\Psi$-ing if the two are related as part to whole and the relata are mutually manipulable. There should be some ideal intervention on $\Phi$ under which $\Psi$ changes, and there should be some ideal intervention on $\Psi$ under which $\Phi$ changes. (Craver 2007: 154)

Thus, the MM account of constitutive relevance can be used to provide a generic demarcation criterion, which “specifies conditions for boundary demarcation independently of special assumptions about the nature of the system whose boundaries are being demarcated.”

Since the MM account of constitutive relevance can be used to tell us what the relevant

---

components of a given mechanism are, it can be used to solve the boundary demarcation problem for cognition and, at the same time, allow for the possibility of extended cognition: Something should count as part of a cognitive process, or mechanism, if it bears a relation of mutual manipulability to the mechanism as a whole. And if the part satisfies the conditions of the MM account yet happens to be located external to the agent’s body, then the hypothesis of extended cognition will have been empirically confirmed. According to Kaplan:
The mutual manipulability criterion is effectively a boundary demarcation criterion, since the total set of mechanism components determined by the criterion marks the inner-outer boundary of that mechanism...it is also a generic demarcation criterion in the relevant sense in that it requires no special assumptions about the nature of cognition. Because intervention-delimited boundaries are resilient to challenges arising from these assumptions, the [debate over extended cognition] can thus be resolved without settling more controversial debates about the nature of cognition. (2012: 557)

Before moving on to our novel argument against EMH, it is worthwhile to point out several challenges facing Kaplan’s proposal even prior to the considerations raised in our argument. The first challenge facing Kaplan’s proposal is that the MM account as it currently stands is incomplete in several respects: First, Craver provides a sufficient condition for constitutive relevance, but explicitly concedes that the sufficient condition may not be necessary for constitutive relevance. Furthermore, Craver models ideal constitutive interventions after ideal etiological interventions, but he only provides an account of an ideal constitutive intervention on the behavior of the parts of the mechanism (with respect to the behavior of the whole)- he fails to provide a corresponding account of an ideal constitutive intervention on the behavior of the whole mechanism with respect to its parts.
The incompleteness may reflect the merely epistemic aspirations in Craver (2007)—namely, providing an account of how we *discover* which parts of a mechanism are relevant components, rather than mere parts. However, if we are to use the MM account to solve the boundary demarcation problem, a merely epistemic version of MM will not suffice, due to the following dilemma: Either a mechanism contains *real* boundaries, or it does not. If it does not contain real boundaries, then there is no fact of the matter as to what counts as part of the mechanism and what doesn’t. And if there is no fact of the matter as to what counts as part of the mechanism and what doesn’t, then there is no fact of the matter as to who is right in the extended cognition debate. Therefore, if a mechanism doesn’t contain real boundaries, then we cannot use the MM account to contribute to the extended cognition debate.

On the other hand, if a mechanism *does* contain real boundaries, then a merely epistemic version of MM cannot tell us what those boundaries are—it can only tell us (some ways in which) we *discover* the boundaries. And if a merely epistemic version of MM cannot tell us what the boundaries are, then it cannot suffice to solve the boundary demarcation problem. The boundary demarcation problem requires a *metaphysical* solution, and therefore the MM account *qua* solution to the boundary demarcation problem must provide a metaphysical account of what it is for a part to be a relevant component.

A second challenge facing Kaplan’s proposal is that, as Bert Leuridan (2012) points out, MM implies that if X’s Φ-ing is constitutively relevant for S’s Ψ-ing, then X’s Φ-ing is a *cause* of S’s Ψ-ing. This is problematic insofar as it is commonly thought there is a strict distinction between causation and constitution. Leuridan suggests that perhaps it would be best to do away with a strict distinction between causation and constitution. However, doing away with the distinction between causation and constitution is problematic in the context of the extended cognition debate.
cognition debate: The less of a difference there is between causation and constitution, there is correspondingly less of a substantive issue of disagreement between those who claim that extracranial objects are components of cognitive mechanisms, and those who claim that extracranial objects are merely causes of the behavior of cognitive mechanisms.

In the rest of this chapter, though, we assume for the sake of argument that Kaplan can find a way to revise his proposal so as to address the challenges above. Our argument against EMH will not depend on any of the particular details of the MM account. Nor will it rely on functional disanalogies between Otto and ordinary human beings that are just a matter of fine detail or degree. Instead, our argument will make use of a necessary condition on holding a belief for a reason, as we begin to discuss in the next section.

4.4. Believing for Reasons

A feature of cognitive states such as beliefs that has not yet been addressed is that much of what we believe is believed at least in part for reasons. Though believing for reasons is not a topic that has been investigated in any great detail by functionalist philosophers of mind, it is also not something that functionalists have any reason to deny is a real phenomenon: if functionalism must deny that we sometimes believe for reasons, then functionalism has a problem that goes far beyond its possible embrace of EMH.95

95 Fred Dretske is perhaps the functionalist (the teleofunctionalist) philosopher of mind who has given the most attention to thinking and acting for reasons. See Dretske (1981) and especially Dretske (1988). Brandom (1994) takes a very different approach that also addresses the question of reasons in the philosophy of mind, and of course Davidson (1980) has a well-known approach to giving reasons a place in the philosophy of mind. Recently, Schroeder (2010) has argued that reductive, naturalistic philosophers of mind need to give more attention than they have to acting for reasons.
Although we do not have a full theory to offer the reader, there is recent independent work by Nomy Arpaly and Ralph Wedgwood suggesting a plausible necessary condition on believing for reasons.\textsuperscript{96} According to these philosophers, to believe something for a reason one must come to believe it in virtue of attitudes that rationalize – that is, stand in epistemically justifying logical or probabilistic relations to – the belief that is formed, and one must come to believe \textit{in virtue of the fact that these attitudes rationalize the new belief}.\textsuperscript{97}

The necessary condition can be seen as stemming from reflection upon the different ways in which a belief can be caused. Some beliefs are caused by merely physical effects, such as banging one’s head against a wall or nefarious surgical intervention: the resulting beliefs are not believed for a reason. Other beliefs are caused in a manner sensitive to (or explanatorily tied to; for present purposes the deep metaphysics of mental causation is not the issue) both physical properties and content, but not to the logical relations between contents. For example, if one is caused, by desiring to believe the best of one’s mother, to believe that she is innocent of the bank robbery for which she was convicted, then one’s belief is caused by (or explanatorily tied to) the content of one’s desire, but again one does not believe for a reason (at least, not an epistemic reason). Finally, some beliefs are caused in a manner sensitive to physical properties, content properties, and the rationalizing relations between the content properties. For example, if one is caused to conclude that Dan will not come to the party by first believing that Dan will


\textsuperscript{97} Thus, our argument presupposes that some beliefs are believed for reasons because of the way in which they are \textit{caused}. The account of higher-level causation developed in Chapter 2 could be used to fully flesh out this account of believing \textit{for a reason}, but it is not required for the general purposes of this chapter.
not come if Deirdre does not and then learning that Deirdre is not coming, with the entailment relation between the premises and the conclusion being part of the cause (or appropriate explanation) of the formation of the conclusion, then one concludes that Dan will not come to the party for a reason.

We adopt the Arpaly and Wedgwood thesis for the remainder of this paper. In part, we direct the reader to these authors’ works for a full defense of the thesis. But we will say a few things here to make it at least prima facie reasonable to adopt the thesis. First, an important point of clarification involving the notion of rationalization. Although Arpaly and Wedgwood hold that having an attitude for a reason requires being caused to have the attitude by appropriate rationalizing attitudes, there is a great deal of latitude available for interpreting “rationalizing” here. Rationalization might well come in degrees, so that one belief might be rationalized by another that provides weak probabilistic evidence in favor of it. Rationalization might well be defeasible, so that a belief that has some weak rationalization might also confront overwhelming evidence that it is nonetheless false. Rationalization might well be independent of the ability to grasp or have correct beliefs about what is rationalized, so that a young child or a dog might have rationalized beliefs. And so on. The heart of the Arpaly and Wedgwood idea is that to believe something for any reason, however poor, one must be caused to believe it in part by the fact that one’s belief stands in some epistemically positive logical or probabilistic relation to at least some of one’s other internal epistemic states. This is a very specific causal relation, but it need not be a very epistemically demanding causal relation.
Additional clarification of the Arpaly and Wedgwood thesis can be provided in response to the following worry\(^98\): Consider Glenda’s belief that 789 times 987 equals 778,743. This belief results from following an algorithm that she was taught in school years ago. Say that she has no idea why this algorithm is a correct one, and say that her belief stands in no epistemically justifying logical or probabilistic relations to the rest of her belief system. Is this a case of a belief that is held for an epistemic reason yet does not satisfy the proposed necessary condition on believing for a reason?

The answer to this question depends on why she uses the multiplication algorithm. Now, on the Arpaly and Wedgwood thesis she needn’t understand why the algorithm is correct in order for her belief to be rationalized. On the other hand, not just any causal process that results in her using the algorithm will likewise result in the belief being rationalized. Cases in which she might use the algorithm, and yet the resulting belief is not held for an epistemic reason, could include using the algorithm as a result of banging her head on the wall, or shuffling a deck of cards with a different multiplication algorithm written on each card, and then randomly drawing one card and using the multiplication algorithm written on that card to arrive at the belief. The more typical causal process, we take it, is one where she has a reason to believe that the multiplication algorithm that she learned in school is the correct algorithm to use, and that is why she uses the algorithm to arrive at the belief. In which case, this does satisfy the proposed necessary condition for holding a belief. Furthermore, in contrast to extended beliefs, Glenda’s belief can be sensitive to her epistemic reasons - for example, she may come to

---

\(^98\) This worry was raised by an anonymous journal referee.
have an epistemic reason to believe that the multiplication algorithm that she learned in school is not the correct algorithm to use.\footnote{For further discussion of this issue see below, 4.6.}

Now, imagine a case parallel to Davidson’s in the theoretical domain. Suppose a particularly dense student shows almost no ability to distinguish between arguments that deny the consequent (modus tollens) and arguments that deny the antecedent, in that the student is vaguely satisfied with the conclusion of either sort of argument so long as the conclusion is not radically inconsistent with something the student already believes (and when it is inconsistent, the student shows no reliable sign of being able to resist the fallacious argument better than the valid argument). And suppose this same student is presented with the theses that, if Pluto is a planet then every Pluto-like heavenly body must also be a planet, yet – because there are hundreds of such bodies within the sun’s gravitational influence and inside the Oort cloud – it would be absurd to hold every Pluto-like heavenly body to be a planet. The student nods sagely and says “so, Pluto just isn’t a planet.” The student has reached this conclusion on the basis of premises he just heard and his inner response to them, but because the student is so insensitive to validity in this domain, it seems much more likely that the student’s belief is formed by something like free association than by the logical relations between the ideas he has just come to believe. (Imagine, not unrealistically, that the student basically remembers the two facts just presented, but if asked to justify the claim that Pluto is not a planet, the student would not reconstruct the \textit{modus tollens} argument but instead say something vaguely incoherent about it being inside the something-cloud and so that meant it isn’t a planet.) And if the student’s belief is formed in this free-associative way, then it seems clear that although the student has a good
reason to believe Pluto is not a planet, the student does not believe it for that good reason. Instead, the student believes for a poor reason (if there is some way in which this sort of free association has positive epistemic status) or for no reason at all (if not).

Moving to a less naturalistic case, imagine that a helpful neurosurgeon is removing a tumor from a patient’s brain when she sees that the patient believes P and believes that P implies Q, but that the patient does not believe Q. Being a helpful neurosurgeon, she adjusts her patient’s brain so as to realize a belief that Q. In this case, as in the more naturalistic one, a person has good reason to believe a thesis, and comes to believe it through a causal pathway that relies on the beliefs in question that make up the good reason, but again in which the belief is not formed in virtue of the good reason. And again, it seems as though the result is that someone believes a thesis, but not for good reasons.

Going further, imagine that the helpful neurosurgeon was given advance warning by the patient herself. The patient told the neurosurgeon “I’m pretty sure I have a problem in my beliefs, with something emotional preventing me from just accepting Q even though everyone knows P and, like most people, I agree that P implies Q. I know that this type of irrationality makes me look foolish, so if you get a chance, could you please revise my belief so that I hold Q?” In this case, as in the previous one, the helpful neurosurgeon can bring the patient’s beliefs into conformity with what she has good reasons to believe, but these adjustments still do not make it true that the patient believes in virtue of the weighty rationalization: she believes Q, but not because of her good epistemic reasons to believe it.

And finally, imagine that the patient is herself a neurosurgeon, and that she performs a surgery on herself to directly manipulate her neurons so as to change her belief that not-Q into a belief that Q. And let us suppose that she does so because she is frustrated with her motivated
irrationality, which is preventing her from believing Q in the ordinary way on the basis of her
good evidence. Still, though she will end up believing Q, and though it will be true that she has
good epistemic reasons to believe it, she will not believe it for those reasons.

Notice that one can agree that in all of these cases the belief held by the protagonist is
simply not caused in the right sort of way, without presupposing that any of the central tenets of
EMH, or of the New Mechanists, are false. Thus, proponents of EMH (including those discussed
in 3.3. who wish to use the New Mechanists’ account of constitutive relevance on behalf of
EMH) shouldn’t object to the proposed necessary condition on believing for reasons. In the next
section we will argue that accepting this necessary condition leads to a difficulty for the
proponent of EMH who wishes to extend the mind to include Otto’s notebook, as well as any
external states that are generated via practical agency.

4.5. A Novel Argument against EMH

The final variant in which the patient directly manipulates her own beliefs is obviously a
close parallel to the story of Otto. If the self-manipulating patient does not believe Q for a
reason, then- even if the marks in Otto’s notebook are beliefs- the marks in Otto’s notebook are
not believed for reasons. They are caused in the wrong way, through a deviant causal chain.

To see the problem for Otto in detail, think again of his case. Otto believes that he is in
front of MoMA, and he believes he is on 53rd street in New York. His beliefs thus license a
conclusion: that MoMA is on 53rd. Suppose that, because of the first two beliefs, Otto writes
“MoMA is on 53rd street in New York” into his notebook. Is the process one in which the
notebook inscription is there because the attitudes that rationalize the inscription (the putative
belief) caused it in virtue of rationalizing it? It seems the answer is ‘no’. Otto’s beliefs, in combination with a desire to remember the location of MoMA, and in combination with his various beliefs about his notebook, his habits regarding the notebook, and so on, combine to cause an intention to write “MoMA is on 53rd street in New York” into the notebook. They rationalize this action, the writing, on the basis of their combined weight as a reason for writing the words “MoMA is on 53rd street in New York” into the notebook. But, just as in the self-manipulating patient case, just because one may perform an action for a reason—such as performing surgery on oneself to change one’s beliefs, or writing down the location of MoMA into one’s notebook—it doesn’t follow that the belief itself is had for a reason. Otto’s epistemic rationalization is not able to non-deviantly cause the inscription in the notebook, because the only mechanism available for inscribing in the notebook is one that works via practical agency.

In contrast, Inga’s belief concerning MoMA’s location is a belief held for a reason. At least, it is if we assume that her belief was caused by, say, her perceiving that MoMA is on 53rd, rather than by banging her head against a wall. Inga’s perception of MoMA’s location (or her belief about her perceptual state) rationalizes her belief concerning MoMA’s location, and her belief concerning MoMA’s location is had in virtue of the rationalizing attitude (her perception of MoMA’s location). Thus, Inga’s belief meets the necessary condition on having a belief for a reason, whereas Otto’s notebook inscriptions do not meet the necessary condition on having a belief for a reason.

The reader might wonder at this point why we are insisting upon a difference between Inga’s intracranial belief and Otto’s extracranial belief, when it seems that both beliefs result from a causal process stemming from a perception of MoMA’s location. However, to ignore the difference between how the two beliefs are caused is to ignore the difference between deviant
and non-deviant causal processes. Inga’s intracranial belief results from a non-deviant causal process stemming from a perception of MoMA’s location: The rationalizing reason for her intracranial belief is also the cause of her intracranial belief. And while it is true that Otto’s extracranial belief likewise results from a causal process stemming from a perception of MoMA’s location, in his case the causal process is deviant: It is not the case that the rationalizing reason for his extracranial belief is also the cause of his intracranial belief. The cause of his extracranial belief is an intentional action—writing down the location because he wants to be able to remember it—thereby severing the direct connection between the rationalizing reason and the notebook inscriptions. To reiterate, it is important to notice that Otto’s case points to an in principle difficulty for the proponent of EMH who wishes to extend the mind to include external states that are generated via intentional action.\footnote{Again, our argument is not directed towards proponents of EMH who wish to extend the mind to include socially distributed cognition.} External states that are generated via intentional action, and not directly in virtue of other attitudes that rationalize them, are, necessarily, disqualified from being counted as beliefs held for reasons, given the necessary condition on holding a belief for a reason.

Our argument begs no questions, for the necessary condition on holding a belief for a reason doesn’t assume that any of the central tenets of EMH, or ordinary functionalism, are false. And, while our argument relies on certain functional details, they are functional details we have a principled reason to rely upon. So our argument is an advance over previous objections to the thesis that Otto exemplifies EMH.
So far, our conclusion is that beliefs had for reasons cannot be extended in the sense specified by EMH. While this conclusion still leaves part of EMH intact, if true it should be quite troubling for proponents of EMH. If the extended mind does not include anything believed for a reason, then the extended mind is also not involved in good (reason-based) reasoning. However, we can also argue for a stronger thesis, that no Otto-like beliefs can be extended, if we help ourselves to the premise that every token belief is something that could have been had for a reason. If this premise is correct, then Otto’s notebook inscriptions cannot count as beliefs, given that it is impossible (as argued above) for them to be beliefs held for reasons.

Just as with the argument for the weaker thesis, the argument for this stronger thesis begs no questions, for the necessary condition on something being a belief doesn’t assume that any of the central tenets of EMH are false. And just as with the argument for the weaker thesis, the crucial premise in the argument for the stronger thesis is one that proponents of EMH have independent reasons to accept. After all, although we once again do not have a full theory of beliefs to offer the reader, it is difficult to think of cases involving mental states that, given how they are causally situated, (a) could not have been had for a reason, yet (b) still qualify as beliefs. And surely anything that meets the minimum requirements (whatever they may be) for being a belief should be able, in principle, to be inferred on the basis of simple rules of inference such as modus ponens. If one cannot, even in principle, infer a “belief” on the basis of modus ponens, then according to the premise above (used in support of the stronger thesis) it isn’t a genuine belief.

101 This is orthogonal to debates between internalists and externalists concerning epistemic justification. Externalists about epistemic justification claim that a token belief can be had for no reason, but this isn’t the same as claiming that a token belief couldn’t be had for a reason.
4.6. Objection and Reply

It might be objected that there exists a closer parallel to the story of Otto than what we have discussed, and moreover that the agent in this case does believe Q for a reason. The problem case we have in mind\(^{102}\) is a case where the self-manipulating agent first comes to believe Q for the appropriate rationalizing reason, but then continues to have the belief in virtue of, say, squirting a certain amount of a particular neurotransmitter, by hand, into her brain. If the agent initially did believe Q for a reason pre-memory enhancement, shouldn’t we say that the agent still believes Q for a reason post-memory enhancement?

One of the nice, yet complicating, features of this example is that it involves focusing on more than just one cause of the agent presently believing Q. Indeed our previous discussion involved a great deal of idealization, for it will rarely (if ever) be the case that a rationalizing reason is the one and only one cause in virtue of which an agent presently believes Q. In general, the rationalizing reason will be one of several causes in virtue of which Q is held. Another ubiquitous cause, for example, will be the presence of a sufficient quantity of nutrients flowing through the agent’s brain. These, of course, are merely physical causes (acting in the background, so to speak), but they are causes nonetheless.

Returning to the case above, we are imagining that the agent starts by believing Q in virtue of having ordinary rationalizing reasons for believing Q, but then continues believing Q

---

\(^{102}\) Suggested by David Chalmers. This case bears a close resemblance to the case of Wanda, who “has a portable electronic memory device that is directly jacked into her brain”, discussed in Weiskopf (2008: 271).
remembers Q is the case) in virtue of applying a memory-enhancing drug. There are two possible interpretations of this case. On one interpretation, the drug enhances or works with other pre-existing causes of belief. On this first interpretation, the agent believes Q in part in virtue of having ordinary rationalizing reasons for doing so, and in part in virtue of the action of the drug. The drug, in this first interpretation, is a causal factor like good nutrition, steady oxygen supply, and the like, in that its action is consistent with the belief that Q continuing to be held in virtue of its causal relations to the rationalizing reasons the agent has to believe Q. On a second, different interpretation, however, the drug is an exclusionary causal factor. We can think of it as some sort of Superdrug: The fact that the agent once believed Q for a reason is irrelevant now that the drug has shown up; after the drug, it is no longer true that the agent presently believes Q in virtue of her rationalizing reasons to do so. Her belief is now utterly insensitive to her rationalizing reasons, and exclusively sensitive to the action of the drug. On this interpretation, the action of the drug is not consistent with the agent continuing to believe Q for a reason. On this second interpretation, the agent is like a person who once checks that the door is locked for a reason, but then acquires a reason-insensitive tic or compulsion of door-checking, and so goes on checking doors in the future for no practical reason at all.

Notice that which interpretation is correct turns on the extent to which the agent’s belief that Q is sensitive to her rationalizing reasons. Thus, if the agent initially believed Q for a reason pre-memory enhancement, the agent still believes Q for a reason post-memory enhancement only to the extent that Q is still sensitive to her rationalizing reasons post-memory enhancement. If, then, this case poses a difficulty for our account, it must be in virtue of illustrating that extended beliefs—such as Otto’s notebook inscriptions—are sensitive to the agent’s rationalizing reasons.
We grant that if it is the case that extended beliefs are sensitive to the agent’s rationalizing reasons, then extended beliefs can qualify as beliefs held for rationalizing reasons. However, we think that the antecedent is false. Consider, once again, the case of Otto. Let us assume that Otto’s initial belief prior to jotting down MoMA’s location in the notebook is had for a reason. The realizer of his belief is a state of his brain, of course. And grant that the notebook inscriptions (partly) constitute a later temporal instantiation of the same belief. Is the belief, at the later time, held for a reason? After all, at the earlier time, the belief is held for a reason, and the existence of the earlier belief is part of how the belief gets consolidated in Otto’s “memory” (in his notebook).

However, as we saw in our discussion above, the mere fact that there is a causal pathway from the earlier belief (held for a reason) to the present belief isn’t enough to guarantee that the present belief is still held for a reason. It all depends on the extent to which the present belief is still sensitive to (in this case) Otto’s rationalizing reasons. And the problem is that the present belief (the notebook inscription) is causally isolated from Otto’s other cognitive states, except via the mediation of Otto’s actions. The notebook inscriptions themselves are not sensitive to Otto’s rationalizing reasons, for the process that generates the notebook inscriptions necessarily involves causation in a way that is independent of, and cut off from, reasons for belief (as opposed to reasons for action). The notebook simply does not have the right sort of causal relationship to Otto’s cognitive states to do better.

Now it is true that Otto could be in an epistemic position such that he could think to himself “if my notebook says X, then it is rational for me to believe X”. Does this show that the

---

103 This was pointed out by an anonymous journal referee.
notebook inscriptions can be sensitive to Otto’s rationalizing reasons? We think that the answer is ‘no’, for notice the following difference between beliefs that result from (typical biological) memories, and Otto’s notebook inscriptions: Say Franklin has a belief that he walked his dog yesterday, and this belief results from a memory he has of walking his dog. This belief is sensitive to his rationalizing reasons, insofar as if he subsequently has reason to believe that the “memory” was really a dream, or hallucination, this rationalizing reason can result in his no longer having the belief that he walked his dog yesterday. In contrast, notebook inscriptions are causally isolated from rationalizing reasons. Say Otto’s notebook contains the words “I walked my dog on June 30”. This inscription is not sensitive to Otto’s rationalizing reasons, for if Otto subsequently has reason to believe, say, that someone else wrote those words in his notebook, this may cause him to no longer have the belief that he walked his dog yesterday, but it will have no effect on the notebook inscriptions themselves—unless, of course, a reason for action (such as a desire to remember and keep track of walking his dog) intervenes.

We have provided an argument against the most interesting claims of EMH that begs no questions and employs only premises that proponents of EMH have independent reasons to accept. Our weaker thesis, that beliefs had for reasons cannot be extended in the sense specified by EMH, makes use of a controversial but independent and plausible necessary condition on having a belief for a reason. Our stronger thesis, that no beliefs can be extended in the sense specified by EMH, makes use of a controversial but independent and plausible necessary condition on something being a belief at all. While the stronger thesis of course deals a more devastating blow to EMH, even if one has reservations about the necessary condition on
something being a belief, the weaker thesis is damaging enough on its own to be cause for concern amongst proponents of EMH.

Furthermore, we can potentially use this argument against EMH to help specify more precisely the scope and limits of functionalism, while warding off Mark Sprevak’s (2009) reductio argument against versions of functionalism that appear to entail an even stronger form of EMH than that proposed by Clark and Chalmers. Versions of functionalism that specify the functional roles of mental states too finely fail to capture the intuition that Martians (or any other creatures that differ from us in terms of fine-grained psychology and/or physiology) could have mental states. But Sprevak is right to insist that the plausibility of functionalism requires a principled upper-bound on the coarseness of the functional roles. Perhaps, then, the best way to set this upper bound, at least when it comes to the functional role of beliefs, is to insist on the necessary condition on having a belief for a reason discussed above. As we saw, Clark and Chalmers’ “modest” version of EMH fails to satisfy this necessary condition, and the radical version discussed by Sprevak (which entails, amongst other things, that “simply by picking up a book, I come to believe everything contained in that book”) would surely fail as well. Functionalists, then, would be well advised to incorporate this necessary condition in their specifications of the functional role of beliefs. This, however, is a topic for another occasion.
Chapter 5: Reframing the Debate Between Dynamicists and the New Mechanists

According to the New Mechanists, good explanations in cognitive science consist in showing how a cognitive phenomenon is produced by a cognitive mechanism, in virtue of the activities and organization of the underlying parts of the mechanism.\textsuperscript{104} In contrast, Dynamicists argue that it is dynamical equations, rather than mechanisms, that are doing the explanatory work in (at least some)\textsuperscript{105} cognitive science explanations.\textsuperscript{106} In this chapter I show that two arguments commonly made in the debate between Dynamicists and the New Mechanists are mistaken. Furthermore, although the debate has heretofore been couched as though it’s a dispute about the norms of explanation, I show that to the extent that the different approaches are incompatible, the dispute really turns on an empirical issue.

In 5.1. I present the history of the debate between Dynamicists and the New Mechanists. In 5.2. I use the interventionist account of causal explanation to show why Dynamicists needn’t claim that the explanatory norms of predictivism are sufficient for good explanation, and in 5.3. I use the mutual manipulability account of constitutive relevance to

\textsuperscript{104} See, e.g. Kaplan (2011); Kaplan & Bechtel (2011); Kaplan & Craver (2011); Piccinini & Craver (2011); Zednik (2011).

\textsuperscript{105} For the rest of this chapter, unless otherwise noted, when I attribute views to Dynamicists about explanations in cognitive science, or about cognitive mechanisms, I am doing so with respect to at least some explanations and mechanisms in cognitive science.

\textsuperscript{106} See, e.g. Chemero & Silberstein (2008); Stepp et al (2011); Anderson et al (2012).
show why even the version of New Mechanism that is least hospitable to Dynamicists can, and should, embrace the explanatory role of abstraction. In 5.4. I show that to the extent that Dynamicists and the New Mechanists are incompatible, the dispute really turns on an empirical issue. If it turns out that Dynamicists are right about the empirical issue, then the New Mechanist approach to explanations in cognitive science must be wrong. If, on the other hand, it turns out that the New Mechanists are right about the empirical issue, I show that it doesn’t follow that the Dynamicist approach to explanations in cognitive science must be wrong.

5.1. The Debate Between Dynamicists and the New Mechanists

According to the ‘once-received’ view of scientific explanation, of which Carl Hempel’s deductive-nomological (D-N) model is a prominent example, successful scientific explanation consists in the subsumption of phenomena under natural laws, by way of deductive argument. The deductive element consists of deducing the explanandum from the explanans, and the nomological element consists of the explanans containing at least one natural law from which the explanandum is derived. On this model, then, scientific explanations are successful when the premises contained in the explanans are true, and the occurrence of the phenomenon (the explanandum) can be derived from the premises contained in the explanans, using traditional rules of inference.

This is no longer the received view of scientific explanations, for two main reasons. First, numerous counterexamples to the sufficiency of the D-N model for successful scientific explanation have been raised: Although we may able to derive the height of a flagpole from the length of the shadow that it casts, the length of the shadow cast by a flagpole does not explain
the flagpole’s height. Although we may be able to derive the occurrence of a storm from a barometric reading, the barometric reading itself does not explain the occurrence of the storm. And although we can surely ‘derive’ the failure of John Jones to become pregnant from the fact that he consumed birth control pills, his consumption of birth control pills does not explain why he failed to become pregnant.

The second main reason for the demise of the D-N model has to do with its ability, or lack thereof, to capture the practices of the “special” sciences—e.g. biology, neuroscience, psychology, economics, etc. What are widely considered to be paradigmatically good explanations in these sciences simply don’t conform to the D-N model. One can of course respond to this objection by arguing that explanations in the special sciences should conform to the D-N model, even if currently they don’t. However, many philosophers of science have argued that rather than there being a nearly ubiquitous flaw in the structure of explanations across the special sciences, it is the D-N model itself which is to blame. Moreover, this objection is very much in line with a dominant theme in contemporary post-Hempelian philosophy of science: The emphasis on tethering philosophical theories to scientific practice, while (unlike the logical empiricists) not restricting, or idealizing, legitimate scientific practice to the practices of fundamental physics. Thus, philosophers of science are now keen to look to the practices of the special sciences when constructing theories of scientific explanation.

In this vein, New Mechanists have pointed out that, at least in many of the life sciences, researchers seem to be looking for mechanisms, rather than laws, in order to provide explanations:

Many of the traditional philosophical ideas about science...were most applicable to domains of classical physics. But, starting in the 1970s and 1980s, certain philosophers who turned their attention to the
biological sciences found that these frameworks did not apply all that well to different biological
domains...in many parts of biology one seems to look in vain for what philosophy has commonly taken to
be the principal explanatory tool of science, that is, laws...If one investigates what biologists and
psychologists seek and treat as sufficient for explanation, it turns out to be mechanisms, not laws.
(Bechtel 2008: 9-10)

One issue of debate amongst the New Mechanists concerns how best to characterize
what a mechanism is, but for the purposes of this chapter we can have in mind something like
Bechtel’s characterization:
A mechanism is a structure performing a function in virtue of its component parts, component operations,
and their organization. The orchestrated functioning of the mechanism is responsible for one or more
phenomena. (Bechtel 2008: 13)

New Mechanists have focused almost exclusively on providing an account of constitutive
mechanistic explanation. Thus, according to the New Mechanists, good scientific explanations
involve a description of how a given mechanism is responsible for a given phenomenon in virtue
of the activities, and organization, of its parts. For example, a good explanation of spatial
memory will involve a description of the spatial memory mechanism that is responsible for the
storage of spatial memories, in virtue of the activities, and organization, of the mechanism’s
components- e.g. hippocampal neurons, synapses, and NMDA receptors. With respect to
cognitive science, then, New Mechanists argue that good explanations should involve a
description of how a given cognitive mechanism is responsible for a given cognitive
phenomenon in virtue of the activities, and organization, of its parts.

In contrast, Dynamicists argue that it is dynamical equations, not mechanisms, that are
doing the explanatory work in cognitive science explanations. Dynamical systems are systems
that change over time, and dynamical models track the relationship between variables in a
dynamical system, using either difference equations (if the system changes in discrete steps) or
differential equations (if the system changes continuously). The general dynamicist idea is that it
is a mistake to model cognition in terms of component parts and operations. Rather, cognitive
systems should be modeled via equations that govern the non-linear nature of interactions in
these systems.

In the debate between Dynamicists and the New Mechanists the most widely discussed
example of a dynamicist explanation is the Haken-Kelso-Bunz [HKB] model of bimanual
coordination. The phenomenon at issue here involves the simultaneous movement of one’s
index fingers. Subjects are instructed to repeatedly move their index fingers side to side in time
with a metronome, and when the metronome speed is increased beyond a certain critical
frequency it turns out that subjects are only able to move their index fingers in phase- that is,
the index fingers move simultaneously towards the body midline, and then simultaneously away
from the midline. Thus, if the finger movements start off at lower speeds in an antiphase
pattern- simultaneous movements to the left or right of the body midline- they suddenly switch
to an in phase pattern at higher speeds, and if the finger movements start off at lower speeds in
phase, they stay in phase at higher speeds. According to Dynamicists, the explanation for the
changes in the phase relationship between the two moving index fingers is given by the
following differential equation [HKB equation]:

$$\phi = -asin\phi - 2bsin2\phi$$

In this equation, $\phi$ is a collective variable- that is, a variable that “summarizes the behavior of
the system’s components” (Chemero 2009: 36)- representing the phase relationship between
the index fingers, while $a$ and $b$ reflect the finger oscillation frequencies.
Before moving on to the arguments in the debate, two caveats are in order with respect to the HKB model. First, although it isn’t clear whether the explanandum at issue is a cognitive phenomenon, I will assume for the sake of argument that it is. Second, while there have been modifications made to the HKB model in the years since the original proposal, these modifications don’t affect any of the issues that I will be discussing, and so for ease of exposition I will be sticking with the original version of the HKB model.

5.2. Dynamicists and Predictivism

New Mechanists argue that to the extent that equations in dynamical systems explanations do not refer to the underlying causal structure of cognitive systems, they are not genuinely explanatory, for much the same reason that the D-N model is no longer considered an adequate account of scientific explanation: There is more to (good) explanation than mere descriptive adequacy and predictive capability. Thus, dynamical systems explanations of cognitive systems may be descriptively accurate, and they may generate a wide array of predictions, but they do not explain how the component parts and operations generate the behavior described by the equations.

Much like the New Mechanists sometimes write as though abstraction is an explanatory vice rather than virtue, Dynamicists do, sometimes, write as though dynamicist explanations only satisfy the explanatory norms of predictivism- that is, descriptive accuracy and predictive strength- , and that satisfying the explanatory norms of predictivism is sufficient for good explanation. For example, Anthony Chemero writes
As Van Gelder (1998) points out, dynamical cognitive science can be seen as an attempt to fulfill Hume’s goal of a scientific psychology similar to Newton’s mechanics- a psychology in which cognition would be explained by mathematical laws... laws that predict the behavior of agents in their environments with great accuracy. (Chemero 2009: 77)

Similarly, according to Stepp et al:

Dynamical explanations do not propose a causal mechanism that is shown to produce the phenomenon in question. Rather, they show that the change over time in a set of magnitudes in the world can be captured by a set of dynamical equations. These equations are law-like, and in some senses dynamical explanations are similar to covering law explanations. That is, dynamical explanations show that particular phenomena could have been predicted, given local conditions and some law-like general principles. (Stepp et al 2011: 432)

In this section I use the interventionist account of causal explanation to show why Dynamicists needn’t claim that the explanatory norms of predictivism are sufficient for good explanation. According to the interventionist account of causal explanation, causal explanations in the special sciences needn't cite exceptionless laws akin to the laws of physics in order to be genuinely explanatory. Rather, we should think of good causal explanations in the special sciences as citing genuine causes featured in relatively invariant generalizations:

A generalization describing a relationship between two or more variables is invariant if it would continue to hold- would remain stable or unchanged- as various other conditions change. (Woodward 2000: 205)

There is no exceptionless law such that anyone who smokes x amount of cigarettes will get lung cancer. Nonetheless, the relationship that holds between smoking and lung cancer holds independent of a wide array of conditions, including (but not limited to) where you happen to be located in the world while smoking, why you started smoking, what kind of job you have, etc.

107 That is, causes as defined by the interventionist account of causation.
Of particular relevance in the context of the debate between Dynamicists and the New Mechanists is the idea that equations found in the special sciences can be interpreted as citing genuine difference-making causes featured in relevantly invariant generalizations. Thus, 

...if we are to interpret \( Y = BX + U \) as describing a causal relationship running from \( X \) to \( Y \), what we must mean is that if we were to intervene to change the value of \( X \), then the value of \( Y \) would change in accord with the equation \( [Y = BX + U] \). (Woodward 2003: 319)

When applied to the HKB model of bimanual coordination, then, we can interpret the equation as describing a causal relationship running from oscillation frequencies to the phase relationship between the index fingers, such that if one were to intervene to change the oscillation frequencies, then the phase relationship would change in accord with the HKB model’s equation. Moreover, so long as the generalization expressed by the equation is relatively invariant- for example, it holds independent of whether one is indoors while moving one’s fingers or outdoors; whether it is a cold day or warm day; etc.- we can think of the HKB model as expressing a good causal explanation rather than being merely descriptively, or predictively, adequate: The HKB model is counterfactual supporting, as it describes what would happen if one were to intervene on the relevant variables.\(^{108}\)

5.3. New Mechanists and the Explanatory Role of Abstraction

\(^{108}\) Although in the text I quote Chemero (2009) as an example of a dynamicist who writes as though dynamicist explanations only satisfy the explanatory norms of predictivism, I should note that more recently (Silberstein & Chemero 2013) he makes a suggestion similar to my argument in this section:

...explanations in systems neuroscience are consistent with manipulationist or interventionist theories of explanation in general. Indeed, not just structural decompositions but also dynamical and graphical explanations can be and often are interventionist explanations. Mechanistic accounts of explanations that focus on localization and decomposition have monopoly on interventionist explanation. There is nothing that says the knobs being tweaked must be structural components; they can also be nomological and graphical features. (Silberstein & Chemero 2013: 969)
Dynamicists have made the following argument against the New Mechanists: Dynamical systems explanations are explanatory (at least in part) in virtue of abstracting away from the details of the underlying causal mechanisms, and the explanatory role of abstraction— that is, the omission of details— cannot be captured by the New Mechanists. Thus:

One key feature of such dynamical explanatory models is that they allow one to abstract away from causal mechanical and aggregate micro-details to predict the qualitative behavior of a class of similar systems...The non-linear dynamics of a cognitive system will be multiply realizable or “mappable” with respect to a wide array of diverse underlying causal mechanical stories about the same processes. (Chemero & Silberstein 2008)

It is clear that Dynamicists consider abstraction to be a theoretical virtue. It is less clear whether Mechanists need deny, or reject, the explanatory role of abstraction. Indeed, Levy & Bechtel (2013) argue that, given the importance of abstraction in a plethora of successful scientific explanations, New Mechanists should incorporate the role of abstraction in mechanistic accounts of explanation. According to Levy & Bechtel, New Mechanists can be divided into those who do already embrace abstraction [Call this the Abstraction Strand] and those who do not [Call this the Specificity Strand]. As Levy & Bechtel note, it isn’t clear that the Specificity Strand would actually reject the explanatory role of abstraction. There is, however, a line of thought amongst proponents of the Specificity Strand that at least suggests a commitment to rejecting the explanatory role of abstraction. Thus, Carl Craver— the most

---

prominent proponent of the Specificity Strand\textsuperscript{110} distinguishes mechanism-sketches from “complete mechanistic models” (2007: 113), with the latter being better than the former insofar as the latter includes details that the former leave out. Moreover, Craver insists on distancing his approach from previous approaches to explanation in cognitive science, such as functional analyses, where the emphasis is on the particular computations, or algorithms, being processed in the brain, while neglecting (or “black-boxing”) the details of how those algorithms are implemented. Thus, when distancing his account from Cummins’s account of functional analysis, he writes:

Cummins’s commitment to functionalism also leads him to a view of explanation that is abstracted away from the details of the mechanism that realize the functions. This abstraction makes Cummins’s account inappropriate as an account of specifically mechanistic explanation. (Craver 2007: 122)

Ironically, though, what has been referred to as Craver’s “main contribution to the literature on mechanistic explanation”\textsuperscript{111}—namely, his account of constitutive relevance—can be used to show why even the Specificity Strand of New Mechanists should embrace the explanatory role of abstraction. Or so I shall argue.

Any given mechanism will contain, at least spatially, a large number of parts. For example, glial cells far outnumber neurons in the brain, and are spatially contained within cognitive mechanisms, yet glial cells are generally thought to be irrelevant when it comes to cognition. The trick, then, for the mechanist, is to provide a principled account of the distinction between the relevant and irrelevant parts a mechanism contains. Craver attempts to provide an

\textsuperscript{110} Craver is also the most prominent proponent of the ontic construal of mechanistic explanation, discussed at length in Chapter 3.

\textsuperscript{111} Levy (2009: 140).
account of constitutive relevance by way of extending the interventionist account of causation, whereby the relevant component parts of a mechanism bear a particular relation to the whole—namely, the relation of mutual manipulability. As we saw in earlier chapters, according to interventionism (very roughly) if X is a cause of Y, then if we ‘wiggle’ X we will see a corresponding effect on Y. And, very roughly, according to the mutual manipulability [MM] account of constitutive relevance, if X is a component of Y, then if we wiggle X we will see a corresponding change in Y, and if we wiggle Y, we will see a corresponding change in X. Again, a bit more formally:

X’s Φ-ing is constitutively relevant to S’s Ψ-ing if the two are related as part to whole and the relata are mutually manipulable. There should be some ideal intervention on Φ under which Ψ changes, and there should be some ideal intervention on Ψ under which Φ changes. (2007: 154)

Craver’s account can be seen as an extension of the interventionist account of causation, insofar as the ideal requisite interventions involved in the mutual manipulability relation are modeled after the ideal requisite interventions involved in the (interventionist) causal relation. Ideal etiological interventions are supposed to be “surgical” in that they involve controlling for possibly confounding effects that would result in our confusing a case of correlation for a case of causation. Thus, if X is a cause of Y, then there must be an ideal intervention on X with respect to Y such that the intervention results in a change in y, if at all, only via the change in X.

Similarly, ideal constitutive interventions are supposed to be surgical in that they involve controlling for possibly confounding effects that would result in our mistaking an irrelevant part for a relevant component. Thus, if X’s Φ-ing is constitutively relevant to S’s Ψ-ing, then there must be an ideal intervention on Φ under which Ψ changes, if at all, only via the change in Φ,
and there must be an ideal intervention on $\Psi$ under which $\Phi$ changes, if at all, only via the change in $\Psi$.\footnote{Actually, Craver doesn’t include the second conjunct in his version of the MM account of constitutive relevance. Rather, as I noted in Chapter 4, he leaves this part of the account incomplete: He specifies the conditions for an ideal intervention on $\Phi$ with respect to $\Psi$, but fails to specify the conditions for an ideal intervention on $\Psi$ with respect to $\Phi$. This will have no bearing on any of the ensuing discussion.} This account is intended to track the experimental practices of cognitive scientists. For example, in lesion experiments there is an intervention on a part of the brain with corresponding changes in task performance (i.e., an ideal intervention on $\Phi$ under which $\Psi$ changes, if at all, only via the change in $\Phi$), and in PET and fMRI studies there is an intervention on task performance, with corresponding changes in the activities of parts of the brain (i.e., and ideal intervention on $\Psi$ under which $\Phi$ changes, if at all, only via the change in $\Psi$).

Now, against the background of providing an account of constitutive relevance, Dynamicists could press the following \textit{reductio} argument against New Mechanists: Clearly, mechanistic explanations in cognitive science that describe the role played by NMDA receptors in the storage of spatial memories go “lower” than explanations that stop when we get to the hippocampus. But why stop with NMDA receptors? After all, NMDA receptors are composed of atoms, which in turn are composed of various sub-atomic particles. Must the ideal mechanistic explanation in cognitive science include details from quantum physics?

Here is how the MM account of constitutive relevance can be used to respond to the argument above. The first thing to notice is that the interventionist account of causation is a \textit{contrastive} account of causation. Thus, ‘$X$ being a cause of $Y$’ is really shorthand for ‘$X$’s being $x_1$ (rather than $x_2$) is a cause of $Y$’s being $y_1$ (rather than $y_2$)’. For example, on the interventionist account, ‘the drug is a cause of recovery’ is shorthand for ‘the presence of the drug (rather than...
the absence of the drug) is a cause of recovery from disease (rather than lack of recovery from disease)’.

Now, interventionism *qua* contrastive account of causation can be used to show how higher-level properties can be difference-makers insofar as they are proportionate to their effects in a way that their underlying lower-level properties are not. Thus, in the case of the spatial memory mechanism, consider the following counterfactuals:

(1) If a given NMDA receptor were to be changed by an intervention from being activated to being inactivated, then a given spatial memory would not have been stored.

(2) If the composition of the given NMDA receptor were to be changed by an intervention from one configuration of subatomic particles to another configuration of subatomic particles, then a given spatial memory would not have been stored.

While (1) is plausibly true in the case where we change the NMDA receptor from being activated \((X = x_1)\) to being inactivated \((X = x_2)\), it is at least plausible to say that (2) is false in a case where we change the underlying configuration of subatomic particles from *one* of the possible realizing states of the activated NMDA receptor \((B = b_1)\) to a *different* possible realizing state of the activated NMDA receptor \((B = b_2)\). This means that it may be the case that there is *no* ideal intervention on the underlying subatomic particles under which the behavior of the spatial memory mechanism changes. In which case, although the subatomic particles may count as *parts* of the spatial memory mechanism, they are not relevant *components*. This, then, is a principled reason for thinking that an ideal mechanistic explanation in cognition need not include details from quantum physics, and it shows why even the Specificity Strand of New

\[113\] The following discussion is very similar to the interventionist proportionality argument discussed in Chapter 2.
Mechanists should embrace the explanatory role of abstraction in constitutive mechanistic explanations.\textsuperscript{114}

I have argued in this section that, appearances to the contrary, New Mechanists need not reject the use of any abstraction in a good explanation. Indeed, embracing the explanatory value of abstraction is not only consistent with other New Mechanist principles—e.g. the MM account of constitutive relevance; it is also advantageous insofar as it offers the New Mechanists a principled way of avoiding the intuitively undesirable view according to which no mechanistic explanation truly explains (or is a good explanation) until it has provided details going all the way down to the level of fundamental physics.

Thus, if the type of abstraction that Dynamicists have in mind when pressing this objection against the New Mechanists is simply an abstraction from some level of detail of the physical composition of the system at issue— that, the kind of abstraction that one can find in systems whose components are multiply realizable— New Mechanists can reply by invoking the MM account of constitutive relevance. However, if the type of abstraction that Dynamicists have in mind is an abstraction from all details of the physical composition of the system at issue, the response discussed in this section will not work: In order to apply the MM account of constitutive relevance, we must be able (at least in principle) to decompose a system into its parts and operations. Abstracting away from all of these details renders the MM account moot.

That said, I take it to still be an advance in the debate to have shown how, and why, New Mechanists need not reject the use of any abstraction in a good explanation. If there is still a

\textsuperscript{114} The issue of whether the interventionist account of causation can indeed be used to show how higher-level properties can be difference-makers is actually more complicated than it may appear in my discussion here. These complications are discussed in Chapter 2. In this chapter I assume that I have successfully shown that interventionism is indeed compatible with higher-level causation.
substantive issue of disagreement here, it cannot be over the explanatory value of abstraction per se, but rather over a particular type of abstraction employed in Dynamicist explanations. This issue will be discussed further in the final section.

5.4. Remaining Issues of Disagreement?

Although the debate between Dynamicists and the New Mechanists has heretofore been couched as though it’s a dispute about the norms of explanation, in this final section I show that to the extent that the different approaches are incompatible, the dispute really turns on an empirical issue. If it turns out that Dynamicists are right about the empirical issue, then the New Mechanist approach to explanations in cognitive science must be wrong. If, on the other hand, it turns out that the New Mechanists are right about the empirical issue, I will show that it doesn’t follow that the Dynamicist approach to explanations in cognitive science must be wrong.

5.4.1. An Empirical Disagreement

The empirical issue between Dynamicists and the New Mechanists revolves around the separability of the causal contributions made by different components of cognitive mechanisms. Dynamicists could argue that cognitive “mechanisms” cannot be decomposed into separable parts- indeed, something like this claim has been made by some Dynamicists, generally using the terminology of interaction-dominant dynamics and component-dominant dynamics. Roughly,

---

if a system contains component-dominant dynamics, then the causal contribution of the
system’s parts are separable, whereas if a system contains interaction-dominant dynamics, then
the causal contribution of the system’s parts are not separable. But if the causal contribution of
the system’s parts is not separable, then- thinking back to the MM account of constitutive
relevance- we can make the argument that at least for some cognitive mechanisms one cannot
explain how they work by explaining how their components work, and thus the New Mechanists
are wrong insofar as there simply are no mechanisms to provide mechanistic explanations with,
or of, in cognitive science. Here is the argument:

(P1) If the New Mechanists are correct, then good explanations in cognitive science will explain
how cognitive mechanisms work in virtue of how their components work.

*Analysis:* As discussed above, this is supposed to be the critical desideratum on good
explanations in cognitive science that sets the New Mechanists apart from their functional
analysis predecessors.

(P2) The most plausible of mechanistic explanation includes the MM account of constitutive
relevance.

*Analysis:* While I haven’t actually made the case here that the most plausible account of
mechanistic explanation includes the MM account of constitutive relevance, I argued earlier that
the MM account of constitutive relevance at least has the potential to assist the New
Mechanists in addressing the explanatory relevance of abstraction. Moreover, the MM account
of constitutive relevance is currently the only account on offer for distinguishing the relevant
components of a mechanism from its mere parts, and it does so by taking into account the
experimental practices of cognitive scientists, while adopting (and extending) the most
prominent account of causation *qua* philosophy of science- interventionism. Thus, there is a lot
to be said for an account of mechanistic explanation that includes the MM account of constitutive relevance.

(P3) If the causal contributions of the parts of a mechanism are inseparable, then none of the parts will meet a necessary condition - according to the MM account of constitutive relevance - on being a relevant component.

Analysis: As discussed above, ideal constitutive interventions involve controlling for possibly confounding effects that would result in our mistaking an irrelevant part for a relevant component. Thus, a necessary condition for an ideal constitutive intervention on φ with respect to ψ is that it must be possible to hold fixed potential confounders (e.g. other components of the mechanism) while intervening on φ. If it is not possible - i.e., if the parts are actually inseparable, at least in terms of their causal contributions to the behavior of the mechanism as a whole - then none of the parts of the mechanism meet the necessary condition for constitutive relevance. That is, the mechanism may contain parts, but none of these parts are relevant components.

(P4) The causal contributions of the parts of cognitive mechanisms are inseparable.

According to some Dynamicists, the equations in dynamical systems explanations of cognitive phenomena indicate that the systems under study contain a particular type of “small-world architecture” whereby the behavior of the system is neither fully regular nor fully random, and the behavior of the parts in the system cannot be fully isolated from one another.116 Of course the claim that the causal contributions of the parts of cognitive mechanisms are not separable is very controversial, and I will not be weighing in on that debate here. Rather, my claim is that if

---

116 For much more discussion see, inter alia, Sporns (2011).
Dynamicists are correct in claiming that the causal contributions of the parts of cognitive mechanisms are not separable, then this spells trouble for the mechanistic approach to explanations in cognitive science.

(P5) One cannot explain how (some) cognitive mechanisms work by explaining how their components work.

**Analysis:** In order to explain how a mechanism works by explaining how its components work, the mechanism must actually contain components. However, according to (P4), the parts of cognitive mechanisms are inseparable. And according to (P3), if the parts are inseparable, then they are not *components*. Therefore, for at least some cognitive mechanisms - i.e. those where the causal contributions of the parts are not separable - one cannot explain how they work by explaining how their components work.

The upshot of this argument, then, is that the cognitive systems at issue simply don’t operate in the way they would need to in order to be targets of mechanistic explanations. The crucial premise in the argument, of course, is (P4)- an empirical claim about the world. Thus, if it turns out that Dynamicists are right about this empirical issue, then the New Mechanist approach to explanations in cognitive science must be wrong. Moreover, it would be disingenuous for New Mechanists to insist that dynamicist explanations of these systems aren’t really explanations, given that it is the nature of the target that is at fault. And yet it seems implausible to try to simply include explanations of these systems within the rubric of an expanded notion of mechanistic explanation, given the tension between the importance of the MM account of constitutive relevance to the New Mechanists on the one hand, and on the other hand the fact that the MM account of constitutive relevance is to blame for the New
Mechanists’ inability to handle these systems within the current rubric of mechanistic explanation.

That said, one problem for this type of Dynamicist approach is that it is much more difficult to make the case that a mechanism cannot, in principle, be decomposed into separable parts, rather than it just being very difficult, in practice, for the mechanism to be decomposed into separable parts. Perhaps this is in part reflected in the fact that Dynamicists themselves fail to distinguish between these two claims. Consider the following statements made by Dynamicists:

..in interaction-dominant dynamics, on the other hand, coordinated processes alter one another’s dynamics and it is difficult, and sometimes impossible, to assign particular roles to particular components. (Anderson et al 2012: 3)

Interaction-dominant dynamics are the basis of interdependence and emergence; interactions among components dominate the intrinsic dynamics of the components themselves...If each component’s dynamics is entangled with the dynamics of every other component, it can become impossible to isolate components and study them separately. (Van Orden et al 2011: 637)

At least prima facie, these statements seem to support the claim that mechanisms governed by interaction-dominant dynamics cannot in principle be decomposed into separable parts. Yet these statements equally well support the claim that mechanisms governed by interaction-dominant dynamics are very difficult to be decomposed into separable parts. Yet it is the former claim (P4) that is required in the argument above. In the rest of this section, then, I show that even if Dynamicists have overstated the claim about the separability of the parts of cognitive mechanisms- that is, even if New Mechanists are right about the empirical issue
discussed above- it doesn’t follow that the Dynamicist approach to explanations in cognitive science must be wrong.

5.4.2. Disagreement about explanation?

Even on the assumption that the empirical disagreement between Dynamicists and the New Mechanists is resolved in favor of the New Mechanists, Dynamicist explanations may still be sufficient for good explanation. To begin with, as we saw in 5.2. Dynamicists can provide explanations that meet the explanatory norms of interventionism. Thus, to the extent that there is a disagreement here, we should see it as a (yet to be addressed) debate between the two currently prominent accounts of explanations amongst philosophers of science- New Mechanists and Interventionists - both of which are keen to look to the practices of the special sciences when constructing theories of scientific explanation.

That said, I don’t think it is clear that there is, or need be, a disagreement here. Rather, we could be explanatory pluralists, with Dynamicists qua Interventionists addressing one type of explanandum, and New Mechanists addressing a different type of explanandum.\textsuperscript{117} Thus, Dynamicists could acknowledge that cognitive mechanisms can (at least in principle) be decomposed into separable parts, but then argue that decomposition isn’t required for good explanations in cognitive science. Consider the following claim:

\textsuperscript{117}Another way to put this point is to say that Dynamicists and the New Mechanists address different types of questions. So long as “different types of questions” refers to questions about different (even if closely related) phenomena, I see no substantive difference between these two different ways of putting the point. Thus, throughout this section I will be concerned with the contrast between questions of the form “what causes Y” and questions of the form “how does X cause Y”. I think of these questions as addressing different (though closely related) phenomena, and thus different types of explananda.
(i) Certain changes in finger oscillation frequencies cause changes in the phase relationship between two moving index fingers.

Now, one could ask the further question:

(ii) Why do certain changes in finger oscillation frequencies cause changes in the phase relationship between the two moving fingers?

New Mechanists assume that one must answer this further question in order to have given an explanation, and I think we can see Dynamicists as challenging this assumption. We have already seen how the interventionist account of causal explanation can be used to show why Dynamicists needn’t claim that the explanatory norms of predictivism are sufficient for good explanation. If that’s the case, why think that (i) must be supplemented with an answer to (ii) in order to constitute a good explanation? As Woodward puts the point (in a slightly different context):

[We must distinguish] between appealing to S to explain M and explaining why Ss cause Ms. If explanation is possible at all, it must sometimes be possible to successfully complete the former activity without completing the latter. (Woodward 2003: 222)

The fact that New Mechanists do make this assumption is mostly clearly seen throughout an influential article by Robert Cummins from (2000) entitled “‘How does it work?’ versus ‘What are the laws?: Two conceptions of psychological explanation.” Here we find a relatively early instance of the objection that the D-N model of scientific explanation is deficient insofar as it fails to track the unimportance of laws in the special sciences- in this case, psychology. Cummins argues that when one does find “laws” in psychology, they are really something that has to be explained, rather than something that actually does the explaining:
Does the law of effect explain why giving a pigeon Pigeon Chow whenever it pecks a key increases the rate of key pecking? Or does it just restate the phenomenon in more general terms? Surely the correct moral to draw here is that the law of effect is an explanandum, not an explanans. In science, when a law is thought of as an explanandum, it is called an ‘effect’...in psychology, such laws as there are are almost always conceived of, and even called, effects. We have the Garcia effect, the spacing effect, the McGurk effect, and many many more. Each of these is a fairly well confirmed law or regularity (or set of them). But no one thinks that the McGurk effect explains the data it subsumes. No one not in the grip of the D-N model would suppose that one could explain why someone hears a consonant like the speaking mouth appears to make by appeal to the McGurk effect. That just is the McGurk effect. (Cummins 2000: 119)

Cummins is here implicitly ruling out the possibility that an “effect” could be both explananda and explanans. Once again, though, we can use the interventionist account of causal explanation to show that effects can themselves be genuine explanantia, insofar as they represent genuinely causal, as opposed to merely correlational, relationships, and correspondingly give rise to genuine control relationships that are absent in merely correlational relationships. As interventionists have pointed out, the reason why we care about causal (as opposed to merely correlational) relationships is that causal relationships are amenable (at least in principle) to being controlled in a way that correlational relationships are not. If smoking is a cause of, and not just correlated with, lung cancer, then one can, for example, minimize the risk of getting lung cancer by not smoking cigarettes. This is in stark contrast with relationships that are clearly merely correlational, such as the nearly perfect correlation of U.S. spending on science, space, and technology with the number of suicides by hanging, strangulation and suffocation. No one thinks that we can actually reduce the number of suicides by hanging, strangulation and suffocation if we reduce the amount of money spent on science, space, and technology. To take a more serious example: The question of whether watching media violence
causes real-world aggression/violence really amounts to the question of whether, at least in principle, a reduction in media violence would cause a corresponding reduction in real-world aggression/violence.

In order to assess the claim that effects in psychology have no explanatory value on their own, I have sampled five examples of effects (including two of the effects listed by Cummins) in a standard psychology textbook\(^\text{118}\), and I have given them interventionist interpretations— that is, I

(a) List the dependent and independent variables at issue;

(b) List a possible confounding variable\(^\text{119}\) such that, if (i) intervening on the independent variable produces a corresponding effect on the dependent variable when the confounding variable is not controlled for, yet (ii) intervening on the independent variable does not produce a corresponding effect on the dependent variable when the confounding variable is controlled for, then what we have is a case where the independent variable is merely correlated with the dependent variable, but the independent variable is not actually a genuine cause of the dependent variable. For example, if it turned out that when we didn’t control for the effects of working in a coal mine increasing the amount of cigarettes smoked increases the likelihood of getting lung cancer, but when we did control for the effects of working in a coal mine then increasing the amount of cigarettes smoked did not increase the likelihood of getting lung cancer.

\(^\text{118}\) Okami (2013).

\(^\text{119}\) Some of the confounding variables will have been mentioned in the textbook as variables that researchers were careful to control for, while others will simply be variables that seem to obviously need to be controlled for, even if this isn’t explicitly mentioned in the textbook.
cancer, then we should say that smoking cigarettes is correlated with, but not a cause of, lung cancer;

(c) Briefly state the causal explanation, whereby the change in the value of the dependent variable is the *explanandum*, and the change in the value of the independent variable is the *explanans*, and I simply assume that if the confounding variable were in fact controlled for, the causal relationship between the independent and dependent variables would still exist; and

(d) Spell out the implicit control relationship revealed by the causal explanation.

Insofar as all of these effects can readily be interpreted as representing genuinely causal, as opposed to merely correlational, relationships, and correspondingly give rise to genuine control relationships that are absent in merely correlational relationships, I take this to be *prima facie* evidence against the claim that effects in psychology have no explanatory value on their own, and *prima facie* evidence against the New Mechanists’ assumption that (i) has no explanatory value on its own, and must be supplemented with an answer to (ii). Here then are the effects and their interventionist interpretations:

(1st Effect) The Garcia effect [AKA conditioned taste aversion]¹²⁰

Independent variable: toxic consequences of food

Dependent variable: avoidance of food

Confounding variable: anorexia

¹²⁰ Okami (2013: 343-4).
Causal explanation: The toxic consequences (e.g. nausea) of eating a particular food cause the individual to avoid the food in the future.

Control: If one wants to stop eating lots of Oreo cookies, one way to do so would be to spike the next batch of Oreos with a nausea-inducing substance. This will help one to avoid, rather than seek out, Oreos in the future.

(2\textsuperscript{nd} Effect) The Law of effect [AKA operant conditioning]\textsuperscript{121}

Independent variable: the consequences of a behavior
Dependent variable: repeating the behavior in the future
Confounding variable: individual’s mood
Causal explanation: If a behavior has positive consequences for an individual, this causes an increase in the likelihood that the behavior will be repeated in the future
Control: This has been “successfully used to treat psychological distress and alter behavior in prisons, hospitals, and other institutions.”

(3\textsuperscript{rd} Effect) Primacy effect\textsuperscript{122}

Independent variable: order of presentation of words to be memorized
Dependent variable: recall ability
Confounding variable: general intelligence
\textsuperscript{121} Ibid. pg. 347
\textsuperscript{122} Ibid. pg. 389
Causal explanation: being presented first causes an increase in the likelihood that the individual will remember that word

Control: If one has a list of words to memorize, some of which being more important (for whatever reason) to memorize than others, the ones that are more important should be placed at the beginning of the list rather than in the middle.

(4th Effect) Partial reinforcement effect

Independent variable: frequency of conditioning
Dependent variable: how long the conditioning lasts for
Confounding variable: general intelligence
Causal explanation: Reinforcing conditioned behaviors periodically is a cause of their enduring for a long time
Control: This is the logic behind payouts for slot machines, and explains why it can be so difficult to stop playing slot machines.

(5th Effect) Bystander effect

Independent variable: Number of people other than oneself who are in a position to offer help to a stranger in an emergency situation
Dependent variable: Likelihood of offering help to a stranger in an emergency situation

\[123\] Ibid. pg. 351
\[124\] Ibid. pp. 746-7
Confounding variable: personality

Causal explanation: If there are a lot of people other than oneself who are in a position to offer help to a stranger in an emergency situation, this decreases the likelihood that one will offer help to a stranger in such a situation.

Control: If you want people to help you in an emergency situation, try to get them to think that they are the only one who can help you.

Of course Cummins is correct in pointing out that there are further questions about these effects. But he has given us no reason to think that the effects themselves are not explanatory, and we have seen how they are all readily interpreted as representing genuinely causal relationships that give rise to genuine control relationships. Thus, we have good reason to think that the effects themselves are, or at least could be, genuinely explanatory. Likewise, the New Mechanists have given us no reason to think that (i) has no explanatory value on its own, and thus Dynamicists can be seen as arguing that (i) does have explanatory value on its own, and needn’t be supplemented with an answer to (ii) in order to be genuinely explanatory.

More generally, then, we can dissolve the debate between Dynamicists and the New Mechanists by being explanatory pluralists when it comes to explanations in cognitive science: Some explanations (e.g. Dynamicist *qua* Interventionists) address the question of what causes Y, while other explanations (e.g. the New Mechanists) address the question of how does X cause Y. There is no reason to think that one type of explanation is superior to the other- they just address different types of *explananda*. And we can likewise use the explanatory pluralism approach to resolve the lingering issue concerning the explanatory role of abstraction (discussed at the end of 5.3.): Dynamicists *qua* Interventionists can abstract from all details of the physical
composition of the system at issue when addressing the question of what causes Y, while the New Mechanists can abstract from some level of detail when addressing the question of how does X cause Y.

While I myself favor the explanatory pluralist approach to explanations in cognitive science, I must acknowledge that the New Mechanists may disagree with the above diagnosis and insist that they have not, in fact, conflated different explananda. Thus, the New Mechanists may argue—perhaps on philosophical grounds, perhaps on sociological grounds—that explanations in cognitive science do, or should, aspire to tell us more than just what causes Y. They (should) aspire to tell us how X causes Y. Rather than there being different types of explananda, then, there is just one type of explanandum, and addressing the question of what causes Y merely gives us (at best) a partial, incomplete approach to explanation.

However, even if New Mechanists adopt this position, closer inspection of the MM account reveals that the New Mechanists should still, by their own lights, accept Dynamicist explanations as genuinely explanatory. Thinking back to the HKB model: Assume for the sake of argument that it is possible, in principle, to separate the causal contributions of the phase relationship and the finger oscillation frequencies. Assume further that (some) surgical interventions on the phase relationship will have an effect on the finger oscillation, and vice versa. Now, granted this ‘decomposition’ of the bimanual coordination system has been found lacking by the New Mechanists. As Kaplan & Craver complain, “…it [the HKB model] clearly was not originally advanced as a description of the neural or biomechanical components responsible for the experimentally observed behavioral dynamics.”

---

125 Kaplan & Craver (2011: 615).
However, the only reason why Dynamicist explanations of tightly coupled systems whose parts are (in principle) separable would fail to pass the necessary conditions imposed by the MM account of constitutive relevance is because they fail to be in a part-whole relationship, such that the part-whole relationship is defined via spatial containment. Notice, though, that the manipulability conditions are the ones that distinguish explanatorily relevant components from explanatorily irrelevant parts. Thus, it seems as though New Mechanists should say that the part-whole relationship fails to convey explanatory relevance on its own. If that's the case, and if tightly coupled separable systems pass the explanatorily relevant portions of the MM account, then New Mechanists should accept these Dynamicist explanations as genuinely explanatory- unless, of course, they either reject the MM account, or tell us why the part-whole relationship should be explanatorily relevant.

Thus, even if the empirical disagreement between Dynamicists and the New Mechanists is resolved in favor of the New Mechanists, there are several theoretical positions that Dynamicists could take that should be compatible with the New Mechanists: First, Dynamicists can be seen as interventionists addressing a different type of explanandum- that is, addressing the question of what causes Y- than the type of explanandum addressed by the New Mechanists, who address the question of how does X cause Y. Second, even if the New Mechanists argue that explanations in cognitive science do, or should, aspire to tell us more than just what causes Y, Dynamicists can point out that Dynamicist explanations of tightly coupled systems can pass the explanatorily relevant portions of the MM account of constitutive relevance, and therefore can- even by the New Mechanists’ lights- tell us more than just what causes Y. I conclude that there is- or need be- far less theoretical disagreement between Dynamicists and the New Mechanists than is suggested by the debate.
References


