Understanding Cognition

by

Gordon J. Steenbergen

Department of Philosophy
Duke University

Date: __________________________

Approved:

____
Owen Flanagan, Supervisor

____
Kevin Hoover

____
Karen Neander

____
Walter Sinnott-Armstrong

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Philosophy
in the Graduate School of Duke University
2015

Abstract

Understanding Cognition

by

Gordon J. Steenbergen

Department of Philosophy
Duke University

Date: ____________________
Approved:

____
Owen Flanagan, Supervisor

____
Kevin Hoover

____
Karen Neander

____
Walter Sinnott-Armstrong

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Philosophy
in the Graduate School of Duke University
2015

Abstract

Cognitive neuroscience is an interdisciplinary enterprise aimed at explaining cognition and behavior. It appears to be succeeding. What accounts for this apparent explanatory success? According to one prominent philosophical thesis, cognitive neuroscience explains by discovering and describing mechanisms. This mechanist thesis is open to at least two interpretations: a strong metaphysical thesis that Carl Craver and David Kaplan defend, and a weaker methodological thesis that William Bechtel defends. I argue that the metaphysical thesis is false and that the methodological thesis is too weak to account for the explanatory promise of cognitive neuroscience. My argument draws support from a representative example of research in this field, namely, the neuroscience of decision-making. The example shows that cognitive neuroscience explains in a variety of ways and that the discovery of mechanisms functions primarily as a way of marshaling evidence in support of the models of cognition that are its principal unit of explanatory significance.

The inadequacy of the mechanist program is symptomatic of an implausible but prominent view of scientific understanding. On this view, scientific understanding consists in an accurate and complete description of certain "objective" explanatory relations, that is, relations that hold independently of facts about human psychology. I trace this view to Carl Hempel's logical empiricist reconceptualization of scientific understanding, which then gets extended in Wesley Salmon's causal-mechanistic approach. I argue that the twin objectivist ideals of accuracy and completeness are neither ends we actually value nor ends we ought to value where scientific understanding is concerned.

The case against objectivism motivates psychologism about understanding, the view that understanding depends on human psychology. I propose and defend a normative psychologistic framework for investigating the nature of understanding in the mind sciences along three empirically-informed dimensions: 1) What are the ends of understanding? 2) What is the nature of the cognitive strategy that we deploy to achieve those ends; and 3) Under what conditions is our deployment of this strategy effective toward achieving those ends? To articulate and defend this view, I build on the work of Elliot Sober to develop a taxonomy of psychologisms about understanding. Epistemological psychologism, a species of naturalism, is the view that justifying claims about understanding requires appealing to what scientists actually do when they seek understanding. Metaphysical psychologism is the view that the truth-makers for claims about understanding include facts about human psychology. I defend both views against objections.

For Mom.

Contents

Acknowledgements

It is an author's privilege to acknowledge those whose support contributed to the completion of a work. And since this dissertation is the most well-considered result of so many years study, there are many people to thank. Thank you to those in the philosophy departments that I was so fortunate to be a part of, and whose mentorship and guidance have made me the philosopher that I am today. At the University of Birmingham, thank you to Iain Law, Harold Noonan, and Joss Walker. At Tufts University, I am especially grateful to Jody Azzouni, Nancy Bauer, Daniel Dennett, David Denby, and George Smith. At Duke University, thank you to my advisor, Owen Flanagan, who first helped me formulate the idea for this project, and who has guided and taught me every step of the way to its completion. I could not have asked for a better advisor. I am also especially grateful to Kevin Hoover, Karen Neander, Alex Rosenberg, and Walter Sinnott-Armstrong. I only hope this thesis does some justice to their generosity and patience. Thank you also to Duke University and the Department of Philosophy for funding the past 7 years of study and research, and to Fred and Barbara Sutherland for their financial support during the Summers of 2011 and 2013.

Being a graduate student at Duke was, for me, a flourishing life. This is largely due to the friendships I made there. Also, philosophy happens in conversation, and I learned so much from my peers. Thank you especially to my officemates Matt Braddock, Nate Gindele, and Steve Martin. Never before have I experienced such

camaraderie, both in philosophical discussion and during the basketball games that kept us sane. Core 4. I am also grateful to Bre Kerchner, Carlos Mariscal, Heather Mayer, and Susanne Meyer.

I started my career as an engineer and made my way incrementally to philosophy. Thank you to Tom Parrish, Alan Lee, and Tim Wierbinski for allowing, even encouraging me to continue part-time work as an engineer even as I pursued very different interests. Without that support it is doubtful that any of this would have been possible.

I am so very grateful to Jack, my father, for being a great dad all around and for persevering without judgment as his son strayed from a well-paying and stable profession to pursue one that is neither of these things. Thank you also to my sisters, Alex and Terri, for being so awesome and for their constant friendship.

There are several who did not live to see this work's completion, but whose influence has been immense. First and foremost is Kate, my mother. She taught me so much and was an amazing mom. She continues to guide me every day. Also Harry, my grandfather, who first inspired in me the pursuit of an intellectual life. And my grandmother, Doris, who supported that pursuit in so many ways. There is Fred Dretske, a remarkable human being, and whose participation in a reading group on mental causation during the Spring 2012 semester was one of the highlights of my intellectual career. And also Iris Einheuser, a brilliant young philosopher whose generous mentorship and clarity of thought have been a constant inspiration.

Finally, thank you to Becky, my love, my wife, and partner in all things. A lot of life happens over the course of seven years when you attend graduate school in your thirties. In our case, the birth of our three amazing girls and doing our best to raise them, the death of my mother, five moves to five cities in three countries and a lot of travel besides. Becky has been at my side through it all and I am so very grateful that is true.

1

Introduction

In his 1892 *Principles of Psychology* William James characterizes the mind science of his day as a "string of raw facts; a little gossip and wrangle about opinions", concluding that, "[t]his is no science, it is only the hope of a science" (James, 2001). Just over a century later, James's hope of a secure and progressive science of the mind/brain appears to have been realized in the interdisciplinary enterprise that is cognitive neuroscience, which promises not merely a "string of raw facts" but to explain cognition and behavior. How does cognitive neuroscience explain? My thesis is concerned with this question. I approach answering it by pursuing two lines of inquiry. The first concerns an investigation into the explanatory practices of cognitive neuroscience, the kinds of models cognitive neuroscientists take to be explanatory, and the evidence they marshal in support of those models. The second concerns philosophical theory about the norms of explanation, a theory that tells us when explanations are good explanations, and that provides reasons for thinking this.

I do not think either line of inquiry has priority over the other. I take seriously the Kuhnian idea that normative philosophical theories are beholden to actual scientific practice (Kuhn, 1996). The reason is also Kuhn's: when they are not, they risk being theories of an enterprise we do not possess, where one of the principal motivations for philosophical attention to the sciences is that they appear to be our most successful epistemological enterprise, and so it is this enterprise that we wish to understand. On the other hand, we are not going to get the norms of explanation from neuroscientists. The role of philosophical theorizing is not merely to bring to the surface general features of a scientific discipline's explanatory practices, but also to produce good reasons for why those practices are actually explanatory, why these practices are successful at generating knowledge about what the world is like in a particular domain. With these points in mind, my approach will be to weigh theoretical considerations against actual practice, while also analyzing and assessing these practices in light of philosophical theories and arguments. By pursuing these lines of inquiry together, I believe we can make significant progress toward answering what our current science of the mind has to say about the nature of cognition, as well as how the sciences generate knowledge about what the world is like.

In Chapters 2 and 3, I critically assess a prominent philosophical theory about the nature of explanation in the mind sciences, according to which cognitive neuroscience explains by discovering and describing mechanisms. I evaluate two interpretations of this thesis: a strong metaphysical thesis (Craver, 2007; Kaplan and Craver, 2011), and a weaker methodological thesis (Bechtel and Abrahamsen, 2005; Bechtel, 2008; Levy and Bechtel, 2013). I argue that on either interpretation, mechanism is vulnerable on normative and descriptive grounds: normative arguments for mechanistic explanation are not persuasive, and mechanism fails to adequately account for what's really going on in the actual practice of this discipline. I defend the latter conclusion by considering research on the neuroscience of decision-making. At least in this research program, cognitive neuroscience inherits the basic methodological assumptions and many of its models from cognitive science, decision-theory and economics, computer science, and cognitive psychology. In important respects, cognitive neuroscience just is cognitive science, as this discipline is traditionally understood (cf. Von Eckardt, 1995), with information from the neurosciences functioning as a rich source of evidence for a variety of models of cognition that are the principal units of explanatory significance in this discipline.

In Chapter 4, I argue that the inadequacy of the mechanist program is symptomatic of an implausible but prominent view of the nature of scientific understanding, according to which genuine understanding amounts to having complete and accurate knowledge of certain objective explanatory relations. I trace the origins of this view to Hempel's logical empiricist account of explanation. Ultimately what I think gets lost in this tradition is the idea that the nature of scientific understanding is essentially linked to the kinds of cognitive creatures that we are, our actual cognitive values and aims, and the remarkable capabilities and frustrating limitations characteristic of our cognitive lives.

In Chapter 5, I outline a naturalistic, but normative approach to investigating the nature of scientific understanding as an alternative to objectivism, and defend this approach against objections. On this approach, scientific understanding of a phenomenon amounts to employing certain cognitive capacities toward achieving the ends of understanding, where these ends are, to a significant extent, but not necessarily exhausted by, pragmatic ends. This approach opens investigation of the ends of understanding, and what cognitive strategies we can employ as effective strategies for achieving those ends, to empirical investigation, broadly construed.

2

Cognitive Neuroscience and the Mechanist Thesis

2.1 Introduction

Cognitive neuroscience is, on the face of it, a successful explanatory enterprise. Consider a brief catalog of recent explanatory hypotheses: The dual-process model of the visual system explains how vision works, as well as Balint's syndrome and certain other visual impairments (Milner and Goodale, 1998); the neurocomputational properties of medial temporal lobe explain recall and familiarity in episodic memory (Norman and O'Reilly, 2003); the reference-dependent encoding of perceptual features explains certain perceptual illusions; visual preference over evolutionary time explains certain visual illusions (Purves and Lotto, 2003); the neuroeconomic model of decision-making (Glimcher, 2011) explains the trembling hand phenomenon (Selten, 1975), framing effects (Kahneman and Tversky, 1984), and addiction (Montague et al., 2004); the function and malfunction of the amygdala explains psychopathy (Blair, 2001).

One prominent thesis in the philosophy of the mind sciences that purports to account for this explanatory success is that cognitive neuroscience explains by discovering and describing *mechanisms*, which is to say (roughly), by discovering and describing how the coordinated activity of parts of the system that is the functioning brain produce the various cognitive phenomena that we wish to explain (Machamer et al., 2000; Bechtel, 2008). Call this the “mechanist thesis”.

To assess whether the mechanist thesis is plausible requires stating what this thesis amounts to, a descriptive task that is not as straightforward as one would hope. In fact, the thesis is ambiguous in several respects. In this essay, I construct a taxonomy of and critically assess the mechanists’ commitments by considering what I take to be two important interpretations of the mechanist thesis: a strong metaphysical thesis that Carl Craver and David Kaplan defend, (Craver, 2006, 2007; Kaplan and Craver, 2011), and a weaker methodological thesis that William Bechtel and his co-authors defend (Bechtel and Abrahamsen, 2005; Bechtel, 2008; Levy and Bechtel, 2013). Each interpretation corresponds to a distinctive philosophical approach to explicating the nature of explanation in the mind sciences. The first involves importing normative criteria of explanation from philosophical theory and assessing the practice of the discipline against these criteria. The second involves attributing a particular explanatory methodology to the aims and practices of cognitive neuroscientists based on close attention to these aims and practices. I claim that arguments for the mechanist thesis are, on both approaches, not decisive in its favor. In particular, explanatory variety in the actual practice of this discipline undermines its plausibility.

2.2 Craver and Kaplan on Mechanistic Explanation

Carl Craver is a prominent defender of mechanistic explanation in the neurosciences (Machamer et al., 2000; Craver, 2003, 2006, 2007; Craver and Kaplan, 2011). In a series of recent papers, Craver and David Kaplan have extended this approach to systems and cognitive neuroscience, and so explicitly endorse the mechanist thesis

(Kaplan, 2011; Kaplan and Bechtel, 2011; Kaplan and Craver, 2011). To appreciate how Craver and Kaplan interpret the content of this thesis, consider the following *Mechanist Argument*:

> (1) Cognitive neuroscience is, to a significant extent, concerned with discovering and describing mechanisms.
>
> (2) Explaining a cognitive phenomenon involves discovering and describing the mechanisms responsible for the cognitive phenomenon to be explained
>
> (3) Therefore, cognitive neuroscience explains cognitive phenomena by discovering and describing mechanisms

The first premise of the Mechanist Argument is justified by attention to the aims and practices of neuroscientists. Reviewing what he takes to be a representative sample of recent neuroscience journal articles, Craver concludes that, according to the neuroscience literature, "the brain is composed of mechanisms."(Craver, 2007, p. 2) The task of neuroscience is to discover and describe these mechanisms. So one commitment of Craver's mechanist program is a descriptive claim concerning the investigatory practices of cognitive neuroscientists:

> (C1) Cognitive neuroscience is, to a significant extent, concerned with discovering and describing mechanisms

But what is a mechanism? And what is involved in describing mechanisms such that their descriptions constitute explanations? If the rationale for the commitment to (C1) is the practice of neuroscience, then one should expect that the aims and practices of neuroscientists, in how they talk about mechanisms and in the way they describe them, bears some resemblance to how philosophers who endorse (C1) understand this commitment. And it is not always clear that this is the case. Indeed, neuroscientists appear to use the term “mechanism” in different ways. Together with the fact of significant disagreement among philosophers concerning what a mechanism is, this suggests that some clarificatory work is required to explicate the content of (C1). This is nothing more than the somewhat obvious point that the mere fact that this term appears with high frequency in the descriptions of neuroscientists themselves is not sufficient justification for the truth of the first premise of the Mechanist Argument; it depends on the intended sense of “mechanism” and what constitutes their description. Kuhn is in the background here: we want an account of the enterprise we have, because it is the (apparent) success of this enterprise that we wish to explain. To assess how Craver interprets this claim, we must turn to his analysis of the nature of explanation.

In *Explaining the Brain*, Craver follows Wesley Salmon in defending what Salmon calls an “ontic” conception of explanation (Salmon, 1984), according to which the explanatory relation is an objective, agent-independent relation between *explanandum* and *explanans*—even if there were no human beings around to explain things, it would still be the case that storms explain drops in air pressure and flag-pole heights explain shadow lengths.[1] As Craver writes of mechanisms:

> There are mechanisms (the objective explanations), and there are their descriptions (explanatory texts). Objective explanations are not texts; they are full-bodied things. They are facts, not representations. They are the kinds of things that are discovered and described. There is no question of objective explanations being ”right” or ”wrong,” or ”good” or ”bad.” They just are. (Craver, 2007, p. 27)

[1] Salmon writes, objecting to psychological accounts of explanation: ”we must surely require that there be some sort of objective relationship between the explanatory facts and the fact-to-be-explained. Even if a person were perfectly content with an ’explanation’ of the occurrence of storms in terms of falling barometric readings, we should still say that the behavior of the barometer fails objectively to explain such facts.” (Salmon, 1984, p. 13)

Of course, human beings do not have unmediated access to the world, and so the ontic theorist will concede that we must represent explanatory relations through descriptions, models, diagrams, or some other representational device. To this extent we "explain" a phenomenon via representations as intermediaries. However, for the ontic theorist, those representations constitute good explanations only to the extent that they represent faithfully, accurately, truthfully the objective explanatory relations that there are. One important consequence of this view is that we can conceive of an ideal explanatory text (Railton, 1978, 1981) that represents "all and only the relevant portions of the causal structure of the world." (Craver, 2007, p. 27) On this view, the more detail one is able to include in an explanatory text about the "relevant portions of the causal structure", the better the explanation will be since it represents more closely the objective explanatory facts. Kaplan and Craver concede that explanatory texts might sometimes be better off abstracting from details in order to satisfy various pragmatic constraints. However, this does not, on their view, make such descriptions better explanations than their ideal counterparts. Rather, it just makes them better suited to achieving certain pragmatic ends—for example, it makes them more tractable given our cognitive limitations, or it makes certain information more salient given our pragmatic aims.

Craver's commitment to an ontologically-oriented conception of explanation informs his analysis of mechanisms and how they ought to be described. On his view, a mechanism is a concrete physical object, composed of "entities" and their "activities", that is individuated relative to the phenomenon that it "exhibits". A mechanism, thus conceived, is the objective explanation of the phenomenon in question. Specifically, if ψ is the behavior to be explained, S is the mechanism that explains the phenomenon, $X_1 \ldots X_m$ are the component entities of S, and $\phi_1 \ldots \phi_n$ are the component activities of $X_1 \ldots X_m$ in S, then "S's ψ-ing is explained by the organization of entities $\{X_1 \ldots X_m\}$ and activities $\{\phi_1 \ldots \phi_n\}$." (Craver, 2007, p. 7) Mechanistic explanation, in the representational or descriptive sense, involves describing, as accurately as possible, in a mechanistic model, how ”constituent entities and activities are organized to exhibit a phenomenon.” (Craver, 2007, p. 122) So, whether Craver's interpretation of (C1) is plausible comes down to whether cognitive neuroscience is concerned, to a significant extent, with the discovery of objects of this kind. Given that cognitive neuroscientists appear to use the term ”mechanism” in different ways, and that there is significant enough disagreement among philosophers as to what a mechanism is, there is at least *prima facie* reason to question this claim.[2]

Even supposing Craver could show that cognitive neuroscience, at its explanatory best, is concerned with discovering mechanisms as he describes them, it does not follow that the discovery of mechanisms serves an explanatory function. To show this, Craver must provide a rationale for the second premise in the Mechanist Argument. On Craver's view, this rationale is that the description of mechanisms is in line with a philosophically defensible explanatory strategy—that is, explaining a phenomenon by describing how it is situated in the “causal structure of the world” (Salmon, 1984). “Explanation”, in the descriptive sense, is a success term: on the ontic view, to explain a phenomenon it is sufficient to accurately describe an objective explanation, which, on a causal theory of explanation, is the network of causal relations that give rise to the phenomenon. Furthermore, recall that on Craver's view, a mechanism is that collection of entities whose activities “exhibit” the *explanandum.* So, if the activities of a mechanism exhibit the *explanandum* because they are what *cause* it, then mechanistic explanations are good explanations precisely because they describe the network of causal relations that give rise to the *explanandum.* Indeed, it is an important component of Craver's view of mechanistic explanation that the “activities” of a mechanism are causal relations. To appreciate Craver's rationale for this constraint, and to see how it carries over to Kaplan and Craver's defense of the

[2] I discuss Bechtel's subtle but important alternative interpretation of “mechanism” below.

mechanist thesis in cognitive neuroscience, a helpful contrast is Robert Cummins' account of functional analysis in psychology.

2.2.1 Contrast with Cummins' Functional Analysis

On Cummins' view, one explains a psychological capacity ψ of a system S by describing the "subcapacities" of ψ (Cummins, 1975). However, Craver argues that a description of the subcapacities of ψ is not sufficient for explaining ψ because mere descriptions of subcapacities cannot distinguish between possible and actual explanations. What is required to make this distinction, he argues, is that the subcapacities of ψ "map" onto component parts of the *mechanism* responsible for S's ψ-ing (Craver, 2007, p. 129). Only then does one satisfy the normative constraints on explanation. Cummins rejects this requirement because on his view, it must be possible that the component subcapacities of ψ are also capacities of S. Craver's reply is worth quoting in full:

> Cummins makes this allowance to accommodate "interpretive explanations," which appeal to the flow of information or to the manipulation of representations in a system. Indeed, Cummins is not primarily interested in constitutive mechanistic explanations, but rather in forms of psychological explanation that are functional and largely independent of the implementing mechanisms. I agree with Cummins that these two varieties of explanation must be kept distinct, especially in discussions of explanation in neuroscience. Lumping both together under the rubric of functional analysis blurs this distinction. So let us make it explicit that functional analysis and mechanistic explanations are distinct in that in mechanistic explanation, S's ψ-ing *is not explained merely by the subcapacities of* ψ*, but by the capacities* $\{\phi_1, \phi_2, \ldots, \phi_n\}$*, of S's component parts* $\{X_1, X_2, \ldots, X_m\}$*.*
>
> The distinction is crucial because how-actually explanations are often

> distinguished from how-possibly explanations on the grounds that the latter appeal to component parts that do not exist and because models of mechanisms are often distinguished from sketches on the grounds that the latter contains black boxes or filler terms that cannot be completed with known parts or activities. (Craver, 2007, p. 129)

While Craver draws a sharp distinction between functional analysis and the "constitutive mechanistic" analysis he favors, he also claims that a Cummins-style functional analysis can be incorporated into the mechanistic framework so long as the subcapacities of ψ in the functional analysis correspond to capacities of the component entities that constitute S. Thus, although in *Explaining the Brain* Craver is focused primarily on explanation in neuroscience (i.e. neurobiology), by providing a means of incorporating functional analysis into the mechanist framework, he opens the door to extending this framework to describing the "activities" of a mechanism in more explicitly psychological terms, for example, as representational capacities. However, there is one important qualification, namely, that the subcapacities that get mapped to capacities of S's components parts are analyzable in terms of activities, i.e. in terms of the causal relations among the component parts of S. As Craver writes, the "causal relationships in mechanisms are not mere capacities in Cummins's sense; they are relationships that are potentially exploitable for purposes of control." (Craver, 2007, p. 134) When this condition holds, the explanation of the psychological capacity ψ is "mechanistic" in the appropriate sense, and so is certified as a legitimate explanation of the phenomenon.

It is precisely by exploiting this strategy that Kaplan and Craver extend Craver's ontologically-oriented account of mechanistic explanation to cognitive neuroscience by defending what they call a "model-to-mechanism mapping" (3M) requirement as a "default assumption" concerning how cognitive neuroscience explains and how it *ought* to explain (Kaplan and Craver, 2011). According to this requirement, the variables of an explanatory model “correspond to components, activities, properties, and organizational features of the target phenomenon” and the “dependencies” between these variables “correspond” to specifically causal dependencies between components of the mechanism in question. Based on this analysis, we can infer the following commitments of Kaplan and Craver’s account of explanation in cognitive neuroscience:

> (C2) The activities that component entities of S engage in are causal relationships that are “potentially exploitable for purposes of control”;
>
> (C3) To explain a cognitive phenomenon it is sufficient to describe the mechanism responsible for the phenomenon.

Are causal-mechanistic explanations of the kind Kaplan and Craver endorse sufficient for explaining cognitive phenomena? Suppose that cognitive neuroscience engages in non-mechanistic explanatory practices. Recent work in the philosophy of science has highlighted non-mechanistic explanatory strategies across a range of disciplines, for example, dynamical explanations (van Gelder, 1995), optimality explanations (Rice, 2013), asymptotic explanations (Batterman, 2002), and mathematical explanations (Batterman, 2010; Lange, 2013). The commitment to (C3) suggests that if any of these strategies are employed in cognitive neuroscience, they could be substituted for mechanistic explanations without explanatory loss—it suggests that mechanistic descriptions are *always* appropriate to our explanatory demands where cognitive phenomena are concerned. This is potentially a quite radical thesis if cognitive neuroscience really does pursue non-mechanistic explanatory strategies, if they do so at the expense of explaining mechanistically, and for other than what has been called “merely pragmatic” reasons. An important question becomes whether this is the case.

In fact, Kaplan and Craver's view is even more radical than (C3) suggests, for they explicitly endorse a strong reading of the second premise of the Mechanist Argument, according to which mechanistic description is both sufficient *and* necessary for explaining cognitive phenomena:

> (C4) To explain a cognitive phenomenon it is *necessary* to describe the mechanism responsible for the phenomenon.

In the next section, I consider arguments in defense of this further constraint.

2.2.2 The Necessity of Mechanistic Descriptions

On a strong reading of the Mechanist Argument, *only* mechanistic models of cognitive phenomena are genuinely explanatory. On this reading, the mechanist thesis amounts to mechanistic imperialism: cognitive neuroscience explains a cognitive capacity when and only when it describes the mechanisms responsible for that capacity. Contrast this with a weaker reading of the mechanist thesis according to which mechanistic models are just one among perhaps several classes or kinds of explanations in cognitive neuroscience. While this weaker reading might initially seem appealing, it leads to problems for the mechanist program that Kaplan and Craver defend.

Recall that the mechanist thesis is supposed to provide an account of the apparent explanatory promise of cognitive neuroscience, and a strong reading of the thesis does this. If the description of mechanisms is both sufficient and necessary for explanation, and cognitive neuroscience is in the business of describing mechanisms, an account of its explanatory success follows straightforwardly. However, it may be the case that *only* the strong reading provides such an account. Consider the possibility that only (C3), that mechanistic description is sufficient for explanation, is defensible. In this case, mechanistic models might be one among perhaps several classes or kinds of explanations that one actually finds in cognitive neuroscience. I have listed several examples of the form such explanations might take. A critical question becomes whether there are legitimate or genuine kinds of non-mechanistic explanation and whether any of these are appropriate to the explanatory demands of cognitive neuroscience as well as implicated in its success. If there are, then this *potentially* has significant ramifications for the theoretical appeal of the mechanist thesis since we would no longer have a complete theory of the explanatory success of the discipline. At best, we would have an account of one contributing factor to this success.

One reply to this worry is that it just isn't much of a problem given the actual facts. While we might be able to show that there are non-mechanistic explanations in cognitive neuroscience, the fact that the discipline is overwhelmingly concerned with discovering mechanisms suggests that any cases of non-mechanistic explanation are merely fringe cases that contribute little to our understanding of the discipline's success. Mechanistic description, in this case, would still be what overwhelmingly accounts for this success. Fair enough. However, if we could show not only that cognitive neuroscience explains non-mechanistically, but also that such explanations are what make cognitive neuroscience *succeed* where explanation is concerned, then this would have two important consequences: it would seriously compromise the theoretical significance of a weak reading of the mechanist thesis; and, supposing that it is in fact the case that cognitive neuroscience is in the business of discovering mechanisms, it would suggest that this activity might serve something other than an explanatory function, that mechanistic description is not always or even primarily what explanation consists in, at least in cognitive neuroscience.

Of course, one way to circumvent this challenge is to defend a strong reading of the second premise of the Mechanist Argument, according to which mechanistic description is both sufficient and necessary for explaining cognitive phenomena. And, as noted, Kaplan and Craver explicitly endorse this reading in their defense of the 3M requirement. Specifically, they supply four justifying reasons for their commitment to (C4): mechanistic models provide information necessary for control; mechanistic models "make sense of scientific-commonsense judgments about the norms of explanation"; mechanistic models answer more explanation-seeking questions than do non-mechanistic explanations; and the history of the physiological sciences is one that involves the steady advance of mechanistic explanation. I will consider each of these reasons in turn.

The significance of control

The connection between explanation and control is an important component of Kaplan and Craver's mechanistic view. To control a system is to causally intervene to change the behavior of that system. And mechanistic models of the kind Kaplan and Craver defend, given that they are models of the causal structure of a system, certainly seem to provide this information. Indeed, to the extent that Craver defends a Woodwardian interventionist account of causation (Woodward, 2003), control is built into his account of mechanistic explanation; on his view, explanation consists in information necessary for control. Furthermore, Kaplan and Craver must have in mind that *only* mechanistic explanation provides this kind of information, for otherwise it would not be a justifying reason for their commitment to (C4). But why think control is a condition on explanation? In fact, the connection between explanation and control gets to the core of the mechanistic approach to explanation and it ties into Kaplan and Craver's claim that mechanistic models "make sense of commonsense-scientific judgments about the norms of explanation."

According to Kaplan and Craver, any contender theory of explanation must provide normative criteria that differentiate *merely* "possible", "phenomenal", or "empirically adequate" models on one hand, from explanatory models on the other. To do this, the theory must supply a criterion for "empirical success" (Giere, 2010), that is, an empirical criterion for establishing whether one has been successful in providing an actual explanation rather than a merely possible or pseudo-explanation. The question becomes what criterion fills this role. One suggestion is that only prediction and control can fill this role.[3] And, as Kaplan and Craver point out, *mere* prediction is not enough where explanation is concerned.[4]

To illustrate this point, consider their claim that merely "empirically adequate" models do not explain. A theory or model is empirically adequate if the statements it makes about all observable phenomena are true—that is, if it "saves the phenomena" (van Fraassen, 1980).[5] Empirically adequate theories are thus, by definition, theories whose predictions are borne out by observation. But this does not make them true. Indeed, multiple, potentially inconsistent theories can satisfy the empirical adequacy criterion; and if they are inconsistent then not all of them can be true. So, given that a requirement of the ontic view is that explanatory descriptions are true or accurate descriptions of mechanisms, it follows that prediction cannot be a sufficient empirical criterion of explanatory success. Furthermore, if mere prediction is not enough, and prediction and control are the only available criteria, then control must be involved. Specifically, the criterion of explanatory success must involve successfully predicted outcomes of causal interventions. Mechanistic models, to the extent that they carry information about the causal structure of the system, provide the information necessary to effect such interventions. Therefore, to explain a phenomenon it is necessary to describe the mechanisms responsible for the phenomenon. This is (C4). This argument is compelling, but ultimately unsound.

First, the argument relies on at least two controversial assumptions: that predic-

[3] For example, see Rosenberg (1996).

[4] This is one of the objections to the deductive-nomological account of explanation, according to which explanation amounts to "rational expectation" of the *explanandum* given the explanans (Hempel, 1965), i.e. the occurrence of the *explanandum* can be *predicted* given the explanans.

[5] More specifically, according to van Fraassen, an empirically adequate theory is one whose models specify features that are isomorphic to all of the observable phenomena (van Fraassen, 1980, p. 45).

tion and control are the only contender criteria of empirical success where explanation is concerned; and that mechanistic models are the only kinds of models that carry the kind of information necessary to effect interventions with predictable outcomes. Consider the second of these assumptions in further detail by way of example. Suppose we have a system that we know was designed to perform a specific function, for example, a facial recognition system that, in the presence of faces, outputs certain information about those faces (e.g. their identity). Given this information, it is a straightforward matter to intervene with predictable outcomes without knowing anything about how the system actually works: we just present faces, the identity of which we know, to the system and observe whether it outputs the correct identity of the presented face. We then hypothesize that what explains why the system does this is that it was designed to perform that particular function. So, if we have information about the design of a system, we can successfully intervene to produce predictable interventions simply by manipulating, not the system, but rather the environment in which the system is operating. This is a case of an explanatory model that allows us to generate evidence, in the form of predictable interventions, that the model represents what's really going on, i.e. that the system really was designed to perform the hypothesized function. Of course, this is a strategy that experimental psychology has adopted since its inception. Mechanistic discovery may provide *better* evidence in some cases, but better evidence is not the same thing as the *only* evidence. This leads to a second important point.

Recall that the question at issue is what makes a model explanatory. Kaplan and Craver cite the "fact" that only mechanistic information can differentiate between explanatory and non-explanatory models as a justifying reason for (C4). But even supposing that mechanistic descriptions are the only descriptions that can differentiate between explanatory and non-explanatory models (which I have argued is controversial), it does not thereby follow that the model that does the explaining is itself mechanistic. Rather, it only follows that mechanistic information is the only source of *evidence* that a model is explanatory. It is conceivable that the discovery of mechanisms functions as a reliable source of evidence for a variety of kinds of explanatory models, not all of which are mechanistic. One very important example concerns information-processing models that incorporate contentful representations: we might have mechanistic evidence that the brain is the kind of system that realizes certain information-processing strategies but still think that our models of those strategies are not themselves mechanistic models. In that case, the discovery of mechanisms would serve to marshal evidence in favor of explanatory models that are not themselves mechanistic. If that is right, then (C4) does not follow from premises concerning the presumed evidential value of predictable interventions. Kaplan and Craver would need to show that the only kinds of models that mechanistic evidence is evidence for are themselves mechanistic models. And this they have not shown.

Common-sense scientific judgments about explanation

Kaplan and Craver argue that mechanistic models make sense of what they call common-sense scientific judgments about explanation in a way that other models or descriptions do not. In this particular case, they contrast the explanatory force of mechanistic model with that of dynamical models, whereby a cognitive process is described as resulting from certain macro-level dynamical properties of the system (Chemero and Silberstein, 2008). On their view, a good theory of explanation ought to account for the fact that scientists routinely make certain kinds of judgments about which models are explanatory and which models are not. These "common-sense scientific judgments" include:

- Distinguishing between spurious correlations and genuine explanations
- Eliminating irrelevant conjunctions by identifying only explanatorily relevant

factors

- Distinguishing between phenomenal and explanatory models, i.e. "between models that merely describe a phenomenon and models that explain it"

- Distinguishing between how-possibly models and how-actually models

On Kaplan and Craver's view, what grounds our commonsense judgments in each of these cases is information about the underlying causes of the phenomenon to be explained, that is, it involves "correctly identifying features of the causal structures that produce, underlie, or maintain the explanandum phenomena." (Kaplan, 2011, p. 607) Now, we might agree that this is what grounds scientific judgements in cases where scientists are seeking specifically causal-mechanistic explanations. But why think these judgments apply in all cases, especially when we consider that, as I will argue, cognitive neuroscientists routinely judge non-mechanistic models to be explanatory? If scientists do make such judgments, then mechanism has a real puzzle on its hands. Either mechanists are forced to admit that commonsense scientific judgments extend to other types of explanation, which counts against mechanistic imperialism and also suggests that an appeal to mechanisms does not capture the basis of these judgments. Or, to the extent that scientists endorse non-mechanistic models as explanatory, they are just wrong, which just seems to undermine any appeal to commonsense scientific judgments as a rationale for mechanism. Instead, what's really doing the work are the philosophical norms imposed from philosophical theory.

Consider the specific case of dynamical models. Kaplan and Craver argue that dynamical models are not explanatory based on their claim that defenders of dynamical models ground the explanatory force of these models on their capacity to make accurate predictions. And philosophers have long since rejected "predictivism" as an adequate criterion of explanatory success. But even supposing defenders of dynamical models ought to be interpreted as endorsing predictivism in defense of the explanatory force of these kinds of models, there is a more plausible alternative available. This is that the explanatory force of dynamical models is grounded, not on their predictive success, but on their identifying actual dynamical properties of a system that are relevant to explaining the behavior of that system. Indeed, let us grant that predictivism is not a criterion on which scientists makes judgments of the kinds that Kaplan and Craver describe. Rather, predictive success ought to be understood as a source of evidence (though not the only source of evidence) that the system really does possess the hypothesized dynamical properties.

Explanatory Power

Kaplan and Craver argue that mechanistic models answer a wider range of explanation-seeking questions about a system than do non-mechanistic models. This amounts to the claim that mechanistic models have greater explanatory power than non-mechanistic models. Why think this is the case? As Kaplan and Craver write:

> This follows from the fact that such models allow one to predict how the system will behave if parts are broken, changed or rearranged and so how the mechanism is likely to behave if it is put in conditions that make a difference to the parts, their properties, or their organization. (Kaplan, 2011, p. 613)

Supplied with a complete "explanatory text" of the kind Kaplan and Craver imagine (a complete description of the causal structure of a system), enough time, and computational resources, one would be in a position to answer pretty much any question about how a system would behave under intervention. However, it only follows that one has thereby explained the system if explanation *consists in* this information, i.e.

information necessary for changing the behavior of the system by breaking, changing, or rearranging its parts. If explanation is not primarily mechanistic explanation, then no matter how many questions one might be able to answer about predictions under intervention, it still might be the case that greater explanatory power can be had elsewhere. The point is that the explanatory power of the mechanist approach Kaplan and Craver defend falls out of their commitment to mechanistic explanation, and not the other way around.

The science of cognition and explanatory progress

Finally, Kaplan and Craver appeal to the history of the biophysical sciences as a source of support for their commitment to (C4). The biophysical sciences, they say, have advanced in making explanations that "approximate the ideals" of mechanistic explanation and that this tradition "should not be discarded lightly" in the pursuit of explanatory progress in cognitive neuroscience. I have two replies.

First, the claim that the biophysical sciences approximate the ideals of mechanistic explanation is a controversial claim. For example, the notion of function appears to be an ineliminable feature of many explanations in these sciences. But according to at least one prominent view about functions (Wright, 1976; Neander, 1991), the function of a system is individuated by features that are external to the biophysical system in question, so these features are not causally relevant (in the mechanistic sense) to the behavior of that system. And given that mechanistic descriptions are, on Kaplan and Craver's view, descriptions of all and only those features that are causally relevant to the behavior of a biophysical system, it follows that functional explanations are not mechanistic explanations. Thus, rather than offering support for (C4), the explanatory practices in the biophysical sciences might actually lead us to question the mechanist thesis outright.

But even supposing it were true that the biophysical sciences were mechanistic through and through (in Craver and Kaplan's sense), the *very question at issue* is whether the mind sciences, i.e. those sciences that attempt to explain mental phenomena, ought to be subsumed as just one more of the biophysical sciences. The history and explanatory traditions in the mind sciences over the past 130 years or so consist in a series of alternating tendencies toward mechanistic and non-mechanistic strategies: from dualism to psychophysics and behaviorism, to identity theory and neuroscience, to functionalism and cognitive psychology. And now cognitive neuroscience. One of the central questions at issue is where cognitive neuroscience falls in this tradition. The only way the supposed mechanistic tradition in the biophysical sciences is support for the mechanist program in cognitive neuroscience is if one is already committed to cognitive neuroscience being just another biophysical science, i.e. if one is already committed to the mechanist program.

Although the nature of explanatory progress in the biophysical sciences does not lend support to the second premise of the Mechanist Argument in the way Kaplan and Craver think it does, it does bring to light an important commitment of the mechanist program that they defend. On their view, since according to the strong reading of the mechanist thesis explanatory progress *consists in* mechanistic discovery, cognitive neuroscience makes explanatory progress insofar as it succeeds in discovering the mechanisms responsible for cognition. This amounts to a claim about the nature of explanatory progress in cognitive neuroscience:

> (C5) Explanatory progress in cognitive neuroscience consists in mechanistic discovery.

Another important question becomes how this commitment compares to the actual history of the development of explanatory models in cognitive neuroscience, and whether cognitive neuroscience progresses by the incremental discovery of the mechanisms responsible for cognition.

2.2.3 Taking Stock

I have articulated five commitments of the strong mechanist program that Kaplan and Craver defend as a normative thesis about how cognitive neuroscience explains. Taken together, these commitments provide an account of how cognitive neuroscience succeeds in explaining cognitive phenomena, an account that is meant to be faithful to the explanatory practices of cognitive neuroscience and consistent with a philosophically defensible set of norms for explanation. However, the philosophical arguments for the mechanistic imperialism Kaplan and Craver defend are not persuasive. And the descriptive resources of this account appear, I suggest, meager in the face of the explanatory variety that is characteristic of the discipline—consider the examples with which I began this chapter. What is needed is further investigation into the kinds of explanatory practices that cognitive neuroscience pursues and how this variety of practices advances our explanatory understanding of cognition. In Chapter 3, I consider what I take to be one important example along these lines, namely, the neuroeconomic model of decision-making. Mechanisms, as Craver and Kaplan understand them, may indeed be an important part of the explanatory story in cognitive neuroscience. But it is not at all clear that they exhaust, or even make a significant contribution to, our understanding of what explanatory knowledge about cognition consists in.

2.3 Bechtel and the mechanist thesis

There are at least two significant contrasts between Bechtel's interpretation of the mechanist thesis on one hand, and Kaplan and Craver's on the other. The first is Bechtel's naturalistic approach to investigating the nature of explanation in the cognitive neurosciences. The second is Bechtel's construal of explanation as an essentially epistemic rather than an ontological notion. Both have significant implications for how we ought to understand Bechtel's interpretation of mechanistic explanation. I therefore turn to a discussion of Bechtel's approach along these lines.

2.3.1 A naturalistic approach to explanation

In discussing Kaplan and Craver's commitment to mechanistic explanation as both sufficient and necessary for explanation in cognitive neuroscience, I raised the possibility that if it could be shown that cognitive neuroscience explains non-mechanistically, that this would be an interesting fact about the science of the mind and the nature of explanation. But how could it be that this discipline achieves explanatory success non-mechanistically if the criteria of success are precisely those provided by our best philosophical theory of explanation? Confronted with an apparent non-mechanistic explanation, it is always open to the defender of the mechanist thesis to say, given the virtues of causal theories of explanation, that non-mechanistic models are either incomplete as explanations (until they are supplemented with a mechanistic description) or simply not explanatory.

This is a powerful rationale for the "traditional approach", according to which one starts with a philosophical theory about the foundations of explanatory knowledge (one that is sensibly motivated by the descriptive character of a discipline), and then assesses how the discipline meets the standards of the theory. In other words, the traditionalist challenge to any would-be alternative account of explanation in cognitive neuroscience in this: What rationally justifiable alternative criterion of success *could there be* for any apparent case of explanation being judged an actual explanation other than its being an instance of a general philosophical theory of explanation, given that success is understood from the perspective of philosophical theory? Kaplan and Craver remind us that a great deal of philosophical work has gone into trying to articulate the conditions for successful explanation in the sciences:

> We begin with a reminder from the past 6 decades of philosophical work on scientific explanation (e.g., Salmon 1989): not all empirically (descriptively or predictively) adequate models explain. Furthermore, the line that demarcates explanations from merely empirically adequate models seems to correspond to whether the model describes the relevant causal structures that produce, underlie, or maintain the *explanandum* phenomenon. (Kaplan and Craver, 2011)

Given that causal theories of explanation are equipped to demarcate theories that explain and those that are merely empirically adequate, they argue that their 3M requirement ought to be the "default assumption" concerning both how cognitive neuroscience explains and how it ought to explain. They then issue the following challenge:

> Like all default stances, 3M is defeasible. However, those who would defease it must articulate why mechanistic styles of explanation are inappropriate, what non-mechanistic form of explanation is to replace it, and the standards by which such explanations are to be judged. (Kaplan and Craver, 2011)

Psychological allure and the say-so of scientists are not, presumably, rationally defensible standards by which to judge explanatory success. Indeed, it is precisely one of the functions of philosophical theory to challenge unwarranted psychological bias and the temptations of the fray—to, if not eliminate "explanations" that don't conform with our theory, at the very least to demand a philosophical rationale for their adequacy. And with good historical reason. Our psychological sense of having explained or of understanding something is notoriously unreliable. With regard to the say-so of scientists, although philosophers are well aware of the cost of disputing apparently successful science (i.e. irrelevance), the cost of joining the crowd in what ultimately turns out to be questionable science is perhaps even greater—if there is to be any rational basis for the belief that science in these cases are instances of genuine explanatory knowledge, we need a rationale.[6] The question is: what is this rationale, or more specifically, what is "empirical success" where explanation in cognitive neuroscience is concerned?

Naturalism contrasts with the traditional approach as an alternative method of inquiry in the philosophy of science and it is the approach Bechtel defends as appropriate to investigating the nature of explanation in cognitive neuroscience. Although what naturalism consists in is a contentious issue, Bechtel adopts what can be described as a modest methodological naturalism that rejects the attempt to establish an independent standpoint from which to certify the norms of scientific rationality, i.e. the project of "first philosophy" (Quine, 1969).[7] Bechtel follows Quine in claiming that this traditional approach faces an insurmountable challenge in attempting to advance "criteria independent of science for evaluating science," namely, "to articulate the resources it will employ in defending its evaluations." (Bechtel, 2008, p. 5) [8] Instead,

> the naturalist proposes that we should examine how scientific inquiry is conducted by actual scientists and in doing so avail ourselves of the resources of science. That is, the philosopher of science would focus on securing data about how scientists work and developing theoretical account that are tested against the data. Although such an approach cannot independently specify the norms for doing science, it can draw upon scientists' own identification of cases that constitute good and bad

[6] As Giere notes, "proponents of the new physics of the 17th century won out not because they explicitly refuted the arguments of the scholastics, but because the *empirical success* of their science rendered the scholastics' arguments irrelevant."(Giere, 2010, p.9, emphasis added)

[7] See Rosenberg (1996) and Flanagan (2006) for discussions of varieties of naturalisms.

[8] Bechtel describes logical analysis and conceptual analysis as two examples of what he takes to be failed attempts at realizing such a program.

> scientific practice and use these to evaluate theories about how science works, as well as to evaluate work within the sciences that are the objects of study. (Bechtel, 2008, p. 7)

To put this in terms pertinent to the present discussion of explanation, naturalism is impressed by the claim that there is no independent standpoint from which to assess the significance of cases that conflict with philosophical theory. The naturalistic approach to providing an account of explanation, therefore, involves "developing theoretical accounts" of how cognitive neuroscience explains and testing these accounts against data concerning how cognitive neuroscience proceeds.[9] While this approach cannot independently specify norms for doing science, this need not lead to skepticism concerning the rationality of science since explanatory success need not be understood in the way the traditional program requires, i.e. as defined entirely by the normative criteria supplied by philosophical theory. Instead, the naturalistic alternative is an empirical one: one achieves rational support for the belief that a model is explanatory by the *empirical success* of the model.[10]

A fundamental question then for the naturalist where explanation is concerned is what constitutes "empirical success". The question becomes especially salient when one considers that the single most repeated criticism from Craver and Kaplan is that, when explanation is the goal, empirically successful, but non-mechanistic models are, at best, incomplete: without mechanistic support or augmentation, empirically successful models are, it is said, *merely* possible, phenomenal, empirically adequate, or predictively successful. In other words, explanation is *special* because explanation

[9] To Bechtel's great credit, he has been meticulous in his efforts to carry out this program by investigating an impressively wide array of examples from the cognitive neurosciences.

[10] Of course, this strategy is anathema to the traditionalist: whether science has actually explained anything is precisely the question at issue, and, so the objection goes, simply describing what scientists do does not answer whether what they do is what they ought to do and whether what they achieve is the truth. But the naturalist has an alternative account of what counts as rational support or warrant for a particular (explanatory) model. And the traditionalist program faces the various problems I have mentioned.

and *mere* empirical success come apart.[11] The challenge then is to articulate a notion of empirical success and a notion of explanation that adequately addresses this criticism. However, what naturalism rejects is that any such notion need be developed antecedently to the investigation of the practice of scientists. Rather, where pragmatic considerations warrant it, naturalism proceeds from success as a working assumption, one that can be revised in the course of inquiry, to investigating what explanatory success consists in and how scientists achieve it.

Of course, Bechtel is a champion of the mechanist program, and so he agrees with Kaplan and Craver that cognitive neuroscience explains by discovering and describing mechanisms; he endorses the mechanist thesis. The question is how he understands the theoretical content of this thesis. For given his naturalistic approach, Bechtel cannot help himself to the normative commitments that ground Craver's mechanistic theory. If Bechtel's account has any theoretical content, then it has to be able to rule something out as non-mechanistic. So one important question going forward it whether his account has the resources to do this.

2.3.2 Bechtel on mechanism and mechanistic explanation

On Bechtel's view, explanation is an essentially epistemic notion. To have explained something is a cognitive achievement: explanations are good when a cognitive agent stands in an appropriate cognitive relation with the world. There are the facts about the world—these consist in the various entities, concrete and abstract, and their relations (causal, mereological, etc.) that some or perhaps all of the sciences endorse, and then there is how we represent and interpret these facts such that we gain explanatory knowledge about them. How then, on this view, should we

[11] This criticism appears repeatedly in the history of science. For example, Newton's *Principia* was criticized for merely "saving the phenomena". In one representative example, an anonymous reviewer of the 1st edition says that Newton had written a lovely book of "mechanics" (i.e. mathematics) and that now all he had to do was write a book about physics (Cohen, 1978).

understand mechanistic explanation?

To describe Bechtel's answer to this question, a useful starting point is to consider what he takes to be the *domain* of mechanistic explanation, namely, those systems that are amenable to what he calls "mechanistic decomposition", i.e. a system that can be decomposed, "physically or conceptually" into "parts" and their "operations" (Bechtel and Abrahamsen, 2005). According to Bechtel and Levy, the demarcation between those systems that are amenable to mechanistic decomposition and those that are not has to do whether a system is internally organized (Levy and Bechtel, 2013). Where the internal parts that comprise a system make a specifiable causal contribution to the phenomenon in question, the system is said to be internally organized and so meets the standards for mechanistic decomposition.

As I understand Bechtel, it is then a *further question* as to what counts as an appropriate way of describing or representing the system such that it provides answers to explanation-seeking questions. Bechtel's naturalistic approach requires that he appeal to the practice of scientists to answer this question. To be sure, given the actual practice of scientists, there is a significant degree of overlap between his view and the ontologically-focused view that Craver and Kaplan defend. As noted, Craver concedes that we gain explanatory knowledge about mechanisms via their representations. And Bechtel, while downplaying the notion that explanatory relations exist in the world, counts among his explanatory repertoire descriptions or models of the physical entities and their relations that compose a mechanism. This is especially true of mechanistic explanations in biology. However, Bechtel also provides several other examples of what he takes to be mechanistic explanations that do not fall neatly into this mold. This flexibility is apparent in how he understands a mechanism as consisting of "parts" that perform "operations", rather than, as Craver and Kaplan understand mechanisms, as "entities" and "activities". For example, the parts of a mechanism can be representations, and the operations can be information-processing operations. As Bechtel writes:

> In most biological disciplines, both the phenomena themselves and the operations proposed to explain them can be adequately characterized as involving physical transformations of material substances....
>
> The performance of a mental activity also involves material changes, notably changes in sodium and potassium concentrations inside and outside neurons, but the characterization of them as mental activities does not focus on these material changes. Rather, it focuses on such questions as how the organism appropriately relates its behavior to features of its distal environment...The focus is not on the material change within the mechanism, but rather on identifying more abstractly those functional parts and operations that are organized such that the mechanism can interact appropriately in its environment. Thus mental mechanisms are ones that can be investigated taking a physical stance (examining neural structures and their operations) but also, distinctively and crucially, taking an information processing stance. (Bechtel, 2008, p. 23)

In taking an information processing stance to explain what Bechtel calls a "mental mechanism", one has to appeal to both causal features of the system and to features of the distal environment, in particular, by distinguishing between the vehicles of representation and their content. He then says:

> What is critical to understanding an information processing mechanism is how content is changed as a result of causal operations that change vehicles, for it is in virtue of content that the mechanism is linked to the states outside the mechanism. (Bechtel, 2008, p. 25)

What this means is that, for Bechtel, part of "describing a mechanism" where cognitive neuroscience is concerned will include, if the discipline adopts an "information processing stance", features of the animal's distal environment. This marks a significant contrast between Bechtel, and Kaplan and Craver. Kaplan and Craver require that the description of a mechanism include all and only those features of the system that are causally relevant to the system's behavior. But one of *the* classic problems in the philosophy of mind is how it could be that the intentional properties of functional states could be causally efficacious if those properties are relational properties between mental states and features external to the system that produces behavior, i.e. outside the head. But if the intentional properties of mental states are not causally efficacious in producing behavior then the ontic account of mechanistic explanation excludes them as explanatory contenders (Dretske, 1989). Therefore, Bechtel's interpretation of what constitutes an appropriate description of a mechanism, insofar as it sanctions intentional descriptions, violates the norms embodied in (C2). He must reject it.

2.3.3 Theoretical implications of Bechtel's naturalism

Bechtel's naturalism prevents him from making imperialist claims of the kind endorsed by Kaplan and Craver (e.g. 3M). He must rely on what the practice of cognitive neuroscience tells him, and it is clear that Bechtel believes the mechanist approach to be defensible across a wide range of cases in the cognitive neurosciences. In this respect, he tends toward a strong reading of the claim that cognitive neuroscience explains by discovering and describing mechanisms. But what distinguishes Bechtel's naturalistic interpretation of the mechanist thesis is its revisability in the face of recalcitrant data from the actual practice of neuroscience. He must reject (C4): recalcitrant data, in the form of apparently non-mechanistic explanations, are not evidence of wrongheaded explanatory practices, but rather disconfirming evidence of a strong reading of the theory.

The same point applies to the sufficiency criterion (C3). If investigation into the practice of cognitive neuroscience shows that neuroscientists, at least in some cases, forgo mechanistic explanation in favor of other forms of explanation, then this would be evidence against the view that mechanistic descriptions are sufficient for explanation. Therefore, Bechtel must also reject (C3). Instead, Bechtel ought to be understood as being committed to the working hypothesis that cognitive neuroscience, across a wide range of cases, adopts a methodological commitment to mechanistic explanation, which amounts to decomposing a system into constituent parts and describing how the organized operations of those parts produces the *explanandum* phenomenon in question.

An important question for Bechtel is what would count as recalcitrant data that would require revision of the mechanist thesis—or at the very least, a limitation in its applicability or strength. If mechanistic explanation only involves describing (from any number of perspectives) the parts and operations of a system, then any description of a system that can be *interpreted* as being composed of parts that engage in operations counts as a mechanistic explanation so long as it satisfies this liberal interpretive constraint. The worry is simply that this constraint is so liberal that the mechanistic thesis has very little theoretical content. That is, as a thesis about what explains cognitive phenomena, the worry is that it just doesn't tell us very much. Instead the real explanatory action is in how the parts and operations of a mechanism are described.

2.4 Conclusion

We have two quite different senses of mechanistic explanation that are best understood as corresponding to the distinction between ontic and epistemic construals of explanation on one hand and to traditional versus naturalistic approaches to investigating the nature of explanation in this discipline on the other. On Kaplan and Craver's view, mechanisms are concrete physical objects composed of entities that stand in causal relationships that jointly produce the *explanandum* phenomenon. Mechanistic explanation, in the representational sense, is then simply the accurate description of the entities and relations that comprise these concrete physical objects and their activities. Another crucial feature of this account is that the description of a mechanism includes all and only those features of the system that are causally relevant to producing the *explanandum* phenomenon. Only in this way can one distinguish between explanatory models and models that are merely phenomenal or empirically adequate.

On Bechtel's view, mechanistic explanation amounts to a methodological approach to explaining a cognitive phenomenon that involves decomposing the system responsible for the phenomenon into parts that perform operations. How operations are interpreted is not limited to their causally relevant features—the form the descriptions of these features take can involve multiple different "perspectives" from which to understand the functioning of a mechanism, to include biophysical features, intentional features, and abstract functional features.

It remains to answer whether either conception appropriately captures the practice of cognitive neuroscience, and if so, what this means for the Mechanist Argument. In the next chapter, I build on my analysis of the mechanists' commitments by assessing the mechanist thesis against a representative example of research in cognitive neuroscience, namely, research on the neuroscience of decision-making.

3

Explanatory Success and the Neuroscience of Decision-making

3.1 Introduction

The thesis that cognitive neuroscience explains by discovering and describing mechanisms has both descriptive and normative force. According to this thesis, the explanatory practices of cognitive neuroscience consist in a disciplinary focus on the discovery and description of mechanisms. And because, at least according to a prominent version of this thesis, explaining a cognitive phenomenon requires describing the parts of the system that are causally relevant to producing the explanandum, mechanistic descriptions satisfy the norms of good explanation. Cognitive neuroscience is thus, on this view, perfectly suited to our explanatory aims where cognition is concerned: explanatory progress is not only coincident with, but consists in, the actual advances in mechanistic discovery that are characteristic of the discipline.

In Chapter 2, I argued that the strong mechanism that Craver and Kaplan defend, as a thesis about explanation, is vulnerable on normative grounds. Even supposing knowledge of mechanisms is necessary to adjudicate among explanatory hypotheses, it does not follow that explanations of cognitive phenomena must therefore consist in mechanistic descriptions. And if cognitive neuroscience constructs non-mechanistic models that it intends as explanatory, then given strong mechanism's normative overreach, mechanism would fall short as an adequate account of the apparent explanatory success of this discipline. Instead, an alternative picture suggests itself: that the disciplinary focus on the discovery of mechanisms functions as a source of evidence that can be marshaled in support of a variety of kinds of explanatory models.

In this chapter, my aim is to defend this alternative view as one that more accurately reflects the explanatory practices in cognitive neuroscience and that has important implications for our understanding of the nature of explanation and of the cognitive lives of animals. Since a lot rides on the claim that cognitive neuroscience actually constructs non-mechanistic explanatory models, my aim is to show that this is indeed the case by considering research on the neuroscience of decision-making. I argue that this family of models (Giere, 2010) is both explanatory and non-mechanistic, and so presents a serious challenge to imperialist and naturalistic mechanist theories. In Section 2, I provide a brief overview of the model I aim to discuss as well as its roots in neoclassical economic theory. In Section 3, I discuss the standards for empirical success for neoclassical theory. In Section 4, I describe the neuroeconomic model of decision making and its historical development in further detail. In Section 5, I argue that this family of models is explanatory but not mechanistic.

3.2 The Neuroeconomic Model of Decision-Making

Two landmark papers, the first by Schultz, Dayan, and Montague (Schultz et al., 1997), the second by Platt and Glimcher (Platt and Glimcher, 1999), transformed research into the nature of human decision-making. The first hypothesizes that dopaminergic neurons in the midbrain comprise a mechanism for generating reward prediction errors, known to be a “a computationally tractable method for tracking changes in value” (Smith and Huettel, 2010). The second hypothesizes that regions of the brain responsible for generating goal-directed saccadic eye movements do so by tracking the expected value of a set of outcomes and recruiting this information to guide behavior. Taken together, these hypotheses are the foundation for an explanatory model of decision-making: animals make decisions by tracking the expected values of outcomes and using this information to execute motor responses. In other words, this research provides the basis for what is ultimately an *economic model* of how the brain guides decisions. As Platt and Glimcher write, that “the variables that have been identified by economists, psychologists and ecologists as important in decision-making are represented in the nervous system,” suggests that such a “decision-theoretic model of the sensory motor process may provide a powerful alternative approach to traditional sensory and motor neurophysiological models of response selection.” (Platt and Glimcher, 1999, p. 237)

It’s a remarkable hypothesis and a central organizing principle of the so-called “neuroeconomic” model of decision-making.[1] I will describe this research in some detail because I think it has philosophical significance with respect to both our understanding of a core cognitive function (decision-making) as well as the nature of explanation. There are at least four reasons why this is true of this research in particular. First, research on decision-making constitutes a research program (Lakatos, 1971) that is broadly representative of the interdisciplinary research that is characteristic of cognitive neuroscience. Indeed, the two principal components of the

[1] There is often confusion outside of cognitive neuroscience about what neuroeconomics is supposed to be, for example, that it is an approach to doing economics using information from the neurosciences. However, while the influence of neuroeconomics on such fields as economics, formal epistemology, and decision theory is no doubt potentially very interesting, my principal concern here is with the neuroeconomic model as a model of how the brain functions to realize a particular species of cognitive behavior, namely, decision-making behavior.

neuroeconomic model, how the brain constructs and stores a representation of value, and how this information is recruited for action, comprise a model that integrates research from a range of disciplines—the neurosciences and psychology, economics, biology, computer science, and engineering, each of which employs a variety of investigative techniques and explanatory strategies. Second, the capacity to make decisions to guide behavior is central to the cognitive lives of animals—it is a paradigm case of specifically cognitive behavior, and it is a core investigatory target of cognitive neuroscience.[2] Third, I think we have good reasons to believe that, according to the dominant views concerning the nature of explanation, the neuroeconomic model of decision-making is a family of models that *successfully* explains this capacity in humans and non-human primates, and it does so in the form of a model that describes how the brain functions to guide choice. Finally, as I will argue, this research is explanatory despite being poorly characterized by the theoretical and historical commitments of mechanistic theories of explanation in the cognitive sciences as I described these in Chapter 2, and so it represents a significant challenge to the view that cognitive neuroscience is explanatory to the extent that it is mechanistic.

Before turning to the neuroeconomic model, I will first discuss the neoclassical economic theory on which the model is based. In addition to providing some background that gives the neuroeconomic model context, this discussion of economic theory will be useful in bringing to the surface several important issues, in particular, the different aims of a theory of decision-making, and how we ought to assess whether any particular model achieves these aims. These considerations will in turn be an important part of the argument for the thesis I aim to defend.

[2] I could just say “behavior”. As Dretske notes, if we’re going to explain behavior, we ought to have some idea of what it is we are trying to explain. And the kinds of behaviors we are interested in explaining (as philosophers) are the kinds that, according to Dretske, have *reasons* as their causes; e.g. hiccups and arm twitches are not examples of *cognitive* behavior, whereas opening the fridge for a beer is (Dretske, 1988).

3.3 Neoclassical Economic Theory

Neoclassical economic models of decision-making construe rational choice behavior according to a minimal set of consistency axioms according to which a rational decision-maker is one who makes consistent choices relative to an ordinal ranking of goods. One of the most significant achievements of neoclassical theory is to demonstrate that an agent who makes choices in accordance with these axioms acts as if he chooses the outcome among his feasible opportunities with the highest expected utility, where expected utility is a function of the utility of an agent's chosen goods and the objective likelihood that the choice will be realized. This neoclassical economic model hypothesizes both a descriptive behavioral model of human decision-making, and a norm of rationality: decision-makers are rational (i.e. they make *good* decisions) insofar as they (1) consistently rank goods or outcomes (i.e. have consistent preferences) and (2) make choices accordingly.[3] The question then becomes whether

[3] Significant advances in decision theory occur with conceptual and mathematical innovations in probability theory in the mid 17th century. In particular, Pascal provides a conceptual framework for understanding choice when he proposes that optimal decisions are those that maximize the expected value of a given choice, where expected value is simply the product of two variable, quantifiable parameters: the likelihood that a choice will produce a given outcome, and the objective value of that outcome. The problem is that this model classifies as irrational actual human decision-making behavior that appears perfectly justified from the point of view of an agent's own preferences rather than objective values. Bernoulli's famous example of this problem, the St. Petersburg Paradox (Bernoulli, 1954; Martin, 2013), is a game with an infinite expected value (so one that, on Pascal's model, one ought to play for any price) but one that, in actual empirical fact, few would be willing to bet very much to play (Hacking, 1980). While Bernoulli subsequently modifies the model to better fit actual human behavior by incorporating a chooser's subjective preferences into the model (specifically, as a logarithmic "utility" function of value), this general approach—constructing a model of rational choice that is both descriptively adequate and prescriptive of optimal choice behavior, runs into a fundamental epistemic problem: there is no independent standpoint from which to assess the significance of deviations from the model, i.e. as counterexamples to the descriptive adequacy of the model, as changes in an agent's preferences, or as instances of irrational choice behavior. A solution to this problem, first proposed by Samuelson (Samuelson, 1938), was to posit a general, but provisional constraint on rationality at the outset: rational decision-makers are those whose choice behavior "reveals" consistent preferences, i.e. a consistent ranking of goods. For example, if S chooses A over B, then it cannot be the case that the agent, if rational, prefers B over A; or if S chooses A over B and B over C then S must also prefer A over C. Von Neumann and Morgenstern then demonstrated that a decision-maker who behaves in accordance with these minimal consistency axioms acts as if he is maximizing utility (von Neumann J. and O., 1944). But what to make of deviations from even these basic axioms? I consider this question in further detail

the model is a good model. And the answer depends to a significant extent on ones aims.

3.3.1 Empirical adequacy of the neoclassical model

If economic prediction is the aim (Friedman, 1953), then empirical adequacy (van Fraassen, 1980) is a suitable measure of empirical success. Whether a model is empirically adequate is a function only of whether the model's empirical consequences are true, i.e. whether its predictions are borne out by observation. Thus whether a model is empirically successful where empirical adequacy is the aim depends only on whether our measurements conform with the model's predictions. For this reason, call empirical adequacy a measure-relative feature of a model, where the measure in this case is predictive success.

The neoclassical model performs reasonably well according to this measure under a wide range of circumstances. For example, from this simple axiomatic model, economists have been able to derive many of the theorems of modern macroeconomic theory, theorems that not only predict economic phenomena, but that also inform policies for economic intervention. Thus, the empirical adequacy of the neoclassical model under the range of circumstances relevant to macroeconomics gives the model significant pragmatic value. However, the neoclassical model's predictions deviate from actual human choice behavior under an array of experimental conditions.[4] Consider: over repeated trials of a simple two-outcome choice under risk, experimental subjects choose randomly when the expected utilities of the two outcomes approach the same value (Selten, 1975; McFadden, 2005); subjects choose randomly or avoid deciding altogether when the size of a choice-set exceeds about eight options (Iyengar

in this Chapter.

[4] As Herbert Simon writes: "As soon as we turn from very broad macroeconomic problems and wish to examine in some detail the behaviors of the individual actors, difficulties begin to arise on all sides." (Simon, 1957, p. 197)

and Lepper, 2000; Glimcher, 2011); subjects' choices are sensitive to whether or not they already own a given good (Kahneman et al., 1990); subjects' choices are sensitive to whether a choice is described as a loss or as a gain, even when the expected utility of the outcome remains constant (Kahneman and Tversky, 1984).

These empirical deviations from the model's predictions show that the neoclassical model is not an empirically adequate model of choice behavior in any global sense. On the other hand, given the model's pragmatic value, perhaps deviations under contrived conditions do not really matter that much. In this case, one might opt for a more sophisticated assessment of the model's empirical success. So long as there is some range of application in which the model has significant pragmatic value, the model is an empirically successful model in every way that matters. However, the worry is that without a theoretical basis for establishing what that range of application is, or if there even is some well-defined range at all, the model's pragmatic value comes into question. In particular, it becomes difficult to project the theory to unobserved cases with confidence.[5] One approach to solving this problem is to modify the theory so as to account for deviations from the original model. And a great deal of theoretical economics has been concerned with doing just this. But if we are going to have any confidence in the range of application of any of our models, we need an answer to the question *why* human behavior deviates from our models when it does. That is, at least to some extent, our understanding of the empirical adequacy of the model will depend on how we *explain* deviations from it.

Questions concerning the empirical adequacy of the neoclassical model are further complicated when one considers the normative aims of the theory. Is the neoclassical model an empirically adequate model of rationality? According to the model, observed behaviors that violate the model's consistency axioms are instances of irra-

[5] Given the complexity of modern economies, no two cases will be alike, and yet we have no theoretical basis for determining whether any given differences are differences that will make a difference. In this way, economics is perhaps more like an historical discipline than a science.

tional decision-making. How can we assess this claim? Although empirical adequacy is an empirical criterion, several philosophers have advanced conceptual arguments in defense of a so-called "principle of charity", according to which interpreting a cognitive agent's patterns of belief requires taking as a basic assumption that the agent is rational. According to one particularly stringent version of this principle, human decision-makers must be interpreted as being fully rational if they are to be interpreted at all (Quine, 1973; Davidson, 1973; Dennett, 1978). On an assumption of full rationality, uncharitable interpretations, those interpretations that assign inconsistent patterns of belief to an agent (i.e. patterns of belief that are not truth-preserving), ought to instead be regarded as bad interpretations. This is because part of being a belief at all is to figure in consistent patterns of interaction with other beliefs; if a pattern of cognitive states cannot be interpreted as being consistent, we have no reason to suppose that those cognitive states even amount to beliefs (Stich, 1990).

However, accepted patterns of belief and their linguistic communication serve purposes other than the communication of truths: patterns of belief might be logically inconsistent but nevertheless play an important social function, so "uncharitable" interpretations are sometimes in order (Thagard and Nisbett, 1983). On the other hand, if we are willing to accept such interpretations, it's not clear why we would interpret those as irrational—rather, the point merely suggests that what patterns of belief we accept as rational is wider in scope than what a criterion based purely on logical consistency would sanction. And that brings us right back to where we started: any interpretation of patterns of cognitive states we are willing to admit as consisting in patterns of belief, whether on logical or social-pragmatic grounds, simply gets incorporated into our notion of what rational cognitive patterns consist in. Our theory of rationality will just consist in those patterns of cognitive states we are willing to interpret as patterns of belief, where the source of our willingness to accept a given pattern of belief is having weighed any given interpretation against our intuitions about the norms for belief. As Jonathan Cohen has argued, on the assumption that our intuitions play this justificatory role, a stringent principle of charity is a foregone conclusion since our normative standards for assessing whether human reasoning has gone well or badly are beholden to our intuitions concerning what counts as rational, the same intuitions that ground normal human reasoning (Cohen, 1981).[6] If we accept the primacy of intuition as a basic methodological assumption, Cohen argues that we are forced into attributing to normally functioning human beings a basic rational competence, one that nevertheless allows for variability in how this competence is expressed in performance.[7] This competence/performance distinction allows for some interpretations of irrational belief, but limits interpretations of irrational behavior to errors in performance due, for example, to interferences (Sober, 1978).

If systematic irrationality were a conceptual impossibility, then deviations in behavior from the neoclassical model's predictions would count against the empirical adequacy of the model. However, we need not accept the methodological assumption that our standards of evaluation must be grounded in intuitions about valid inference, and indeed I think we open up a much more interesting conceptual space by sidestepping this assumption and going instrumental, i.e. by evaluating the rationality of patterns of cognitive states according to whether they constitute effective strategies for achieving certain desirable goals.[8] On this view, the question then becomes

[6] Stich makes the central point nicely: "empirical theory of inferential competence must inevitably coincide with the normative theory of inference" since they both rely on the same empirical data, namely, intuitions about the validity of certain kinds of inferences (Stich 1981, p.353).

[7] The competence / performance distinction is analogous to Chomskian arguments concerning a native speakers competence for grammar. Sober 1978 also draws this analogy in the case of inference.

[8] Stich argues that even those patterns of cognitive states that are legitimately interpreted as patters of *belief* (i.e. states that can be given intentional ascriptions because, e.g., as Stich proposes, they track close enough our own hypothetical cognitive states), which are effective strategies

whether the neoclassical model succeeds in its prediction that the decision-procedure it describes is an effective strategy of achieving desirable outcomes—that is, we get a criterion of success that is subject to empirical test.

What complicates an instrumental analysis of rational choice is that both what constitutes a "desirable outcome" and an "effective strategy" can be interpreted in different ways. Human beings value a variety of ends: perhaps the true, the good, and the beautiful among them, but sometimes also ends that conflict with these, even morally reprehensible ends. This suggests we can rightfully make judgments about whether people ought to value the ends that they do, that some ends are desirable for their own sake, independently of whether any particular person values it or not or pursues ends that are in direct conflict with those ends. In that case, we would rate other goods relative to whether pursuing them is an effective means of achieving the intrinsically valuable ends. This in turn leads to the question how to decide what makes an effective strategy. Are strategies effective relative to particular decisions or to a wide range of decision-making contexts? Can a strategy be effective *per se*, and so a strategy for all occasions? Or are strategies only effective relative to alternatives? Are they effective in all possible worlds? Or only in this world, relative to the constraints that this world imposes?

I will set aside the question whether desirable outcomes are the outcomes that individual persons desire or outcomes that are desirable for their own sake. But one way to restrict the range of interpretations of what constitutes an effective strategy is by making the plausible assumption that whether a cognitive agent ought to have pursued a strategy at least depends on whether the agent's psycho-biological constraints permit that strategy as a live option, i.e. that ought implies can. On this assumption, one assesses the effectiveness of a decision-making strategy relative

represent only a small local maximum, and a highly idiosyncratic one at that, in a vast field of possible computational strategies (Stich 1990).

to those constraints. This requires both knowledge of the constraints and how our decision-making strategies are equipped to accommodate them. In this way, empirical considerations inform our normative judgements about rational choice, and our conception of rationality becomes "bounded" by considerations of psycho-biological realism (Simon, 1957).[9] In the specific case of the neoclassical model, given that actual human behavior deviates from the model's predictions, we can ask, when this occurs, whether that is because the agent's cognitive system employs a less effective strategy (compared to the neoclassical model), given the constraints, for achieving the desired outcome, or whether, in fact, the strategy it employs really is effective when those constraints are taken into consideration. Thus, once again, the question as to the empirical adequacy of the neoclassical model turns on how we answer certain explanation-seeking questions about the cognitive system that guides human choice behavior.

I began this section by asking whether the neoclassical model is an empirically successful model, the answer to which depends on our aims. Where a sophisticated empirical adequacy is the aim, I have argued that, in important respects, the jury is still out: in both the descriptive and normative cases, we need to know *why* actual human choice behavior deviates from the model's predictions. Interestingly, where our aims are "mere" predictive success, it turns out we need explanation after all; we need to know whether the neoclassical model *explains* human choice behavior. Although the neoclassical model does not, by itself, have the resources to answer why its predictions deviate from actual human choice behavior, it still makes sense to ask whether the model is explanatory of this behavior. Indeed, according to the neuroeconomic model of decision-making, something like the neoclassical model

[9] Simon advanced the "bounded rationality" view on consideration of models of human choice behavior as "satisficers" rather than utility maximizers. On a satisficing model, a decision-maker chooses a satisfactory, rather than optimal, solution to a given problem. Furthermore, he is rational to do so.

really is explanatory. I therefore turn to a discussion of the neoclassical model's explanatory success.

3.3.2 Evaluating the explanatory success of the neoclassical model

How can we assess the explanatory success of the neoclassical model? In Chapter 2, I produced a set of arguments that aimed to show that mechanistic descriptions are neither necessary nor sufficient for explaining cognitive phenomena. In this chapter my aim is to show that mechanism is vulnerable not only on normative grounds, but descriptive grounds as well. If that is right, if our best current science of the mind, which appears to have significant explanatory promise, is not itself committed to generating only mechanistic explanations, and indeed is actively pursuing non-mechanistic explanations, then given the arguments against mechanism's normative overreach, mechanism falls short as an adequate theory of explanation in this discipline. But to make this objection stick, I need to defend the antecedent to this conditional, that it really is the case that this discpline is engaged in non-mechanistic explanatory practices, that is, I need to show that cognitive neuroscience is committed to models that are both explanatory on the one hand and non-mechanistic on the other. In this section, I consider the first of these issues by considering how we ought to assess the explanatory success of a model. One way to answer this question is to defend a particular normative theory of explanation that provides norms for good explanation. This is the approach that Kaplan and Craver take in their defense of strong mechanism. Rather than attempting the same, my strategy will be to motivate a minimal criterion of explanatory success that allows for non-mechanistic explanations, but by building on what I take to be mechanism's own implicit commitments regarding the norms of explanation. In doing so, my aim will be to show that non-mechanistic models in the neurosciences satisfy this minimal criterion and so, by the mechanists own lights, ought to at least be contenders for explanatorily successful models.

As I discussed in Chapter 2, Craver and Kaplan defend their theory of explanation (and reject others) by assessing normative theories of explanation against what they call "common sense scientific judgments" about the norms of explanation. For example, it is on this basis that they reject the view that dynamical models have explanatory force. I argued that mechanism cannot rely on such judgments to defend mechanistic imperialism if such judgments are routinely made in non-mechanistic contexts. If that is right (and I aim to show that it is), at best Craver and Kaplan's argument shows that the mechanistic theory they defend provides an account of the commonsense scientific judgments that scientists make when they seek causal-mechanistic explanations. More generally, what judgments scientists (and philosophers) make as to whether a model is explanatory will depend on the kind of explanation they are seeking: while identifying underlying causal facts is relevant to producing good causal-mechanistic explanations, identifying underlying statistical facts is relevant to producing good statistical explanations, identifying facts about the content of intentional states (if such facts there are) is relevant to good explanations of behavior, and so on. But if mechanism cannot, by itself, make sense of commonsense scientific judgments across different explanatory contexts, the natural question becomes whether there is some more general commitment that cuts across these contexts that, at least in part, grounds commonsense scientific judgements about explanatory norms. It seems to me that there is. This is the intuition, shared by many (but by no means all), that good explanatory models, at least in part, identify *actual* features of the system that are relevant to explaining the phenomenon in question; that is, the model has to be a true description or accurate representation (or something close enough) of some actual entity or property of the system that is relevant to satisfying ones explanatory aims. In other words, at least one possibility is that common sense judgments about the norms of explanation are grounded by an implicit commitment to realism; good explanatory models are models that tell us something about the way the world really is.[10]

A fruitful way to explore the ramifications of a commitment to realism in explanatory judgments is by considering the philosophical debate about a particular type of reasoning, namely, inference to the best explanation. As this type of reasoning applies to the sciences, the idea is that one infers the truth of a theory or model based on the conclusion that the theory or model is what best explains a particular phenomenon. Nancy Cartwright construes this type of reasoning as taking the following basic argument form (Cartwright, 1983):

$$\begin{array}{l} \quad Q \\ \quad P\ explains\ Q \\ \hline \therefore P \end{array}$$

To secure the validity of this argument, we need to make an additional assumption about what it is for P to explain Q. As Cartwright herself notes, "Many arguments wear their validity on their sleeve: 'I think. therefore, I exist.' But not, 'P explains Q, Q is true. Therefore P is true.'"(Cartwright, 1983, p. 89) This assumption is

[10] Kaplan and Craver certainly appear to endorse such a commitment: "to explain the phenomenon, the model must...reveal the causal structure of the mechanism" (Kaplan and Craver, 2011, p. 605). In discussing abstraction and idealization, they argue that "practices of abstraction and idealization sit comfortably with the realist objectives of a mechanistic science."(Kaplan and Craver, 2011, p. 610) Furthermore, Salmon also clearly shares this ontological enthusiasm when it comes to causal explanation (Salmon, 1984). Michael Strevens writes: "no causal account of explanation—certainly not the kairetic account—allows non-veridical models to explain" (Strevens, 2008, p. 320). Also, Ruben appeals to foundational metaphysical commitments in an argument for why we are justified in thinking certain relations are the explanatory ones (Ruben, 1990). Van Eckhardt and Poland write:

> What is required for an actual explanation is not only a functional analysis but some reason to believe that the system or organism in question actually has (or exercises) C in virtue of actually having (and exercising) the constituent capacities of the functional analysis in the order specified by the functional analysis, namely, the functional analysis must be structurally adequate. (Von Eckardt and Poland, 2004, p. 975)

And it's not just philosophers who share this commitment. Paul Glimcher, one of the founders of the neuroeconomic model that I describe below, in evaluating whether the neoclassical model is an explanatory theory of human choice behavior appeals directly to whether economic variables are actually represented in the nervous system.

the realist assumption that, for something to be an explanation at all, the explanans must be true. On this assumption, "explanations of necessity reflect the structure of the world; it is the nature of explanation that P does not explain Q unless P is true" (Hitchcock, 1992, p. 154).[11] Van Fraassen attacks inference to the best explanation on precisely these grounds: if the best a theory can hope for is empirical adequacy, then we have no grounds for thinking that the theory is true. Furthermore, there doesn't seem to be anything special about the explanation relation that *compels* our belief in the explanans—as Cartwright notes, two-place relations do not usually have this remarkable feature. Absent any warrant for that assumption, goes the anti-realist argument, inference to the best explanation fails as a valid form of reasoning.

However, there is, according to Cartwright, one important exception: if we accept that P *causes* Q, for example, based on "direct experimental testing", we are compelled to believe that P is true, because we are compelled to believe in the existence of the hypothesized entities that figure in the causal chain whose effect is Q. Though Cartwright agrees that inference to the best explanation generally falls to van Fraassen's anti-realism, on her view, "inference to the most probable cause" does not. As Cartwright writes, "What is special about explanation by theoretical entity is that it is causal explanation, and existence is an internal characteristic of causal claims. There is nothing similar for theoretical laws." (Cartwright, 1983, p. 93)[12]

Notice, however, there is something odd about about this line of reasoning. As Hitchcock points out, if we already have independent reason to think that P causes Q, then according to Carwright, we are already committed to the truth of P. So it's

[11] Hitchcock notes that an alternative assumption is that "the world conspires to structure itself so as to reflect our explanatory relations." But Hitchcock, following van Fraassen (van Fraassen, 1980) rejects this as implausible.

[12] To illustrate this point, Cartwright asks us to imagine God telling us each of two explanatory statements: 1) "Schroedinger's equation provides a completely satisfactory derivation of the phenomenological law of radioactive decay." 2) "The rotting of the roots is the cause of the yellowing of the leaves". If we accept that God is telling us the truth, she claims, then we are compelled to believe in the truth of "the rotting of the roots", but not the truth of Schroedinger's equation.

not as if we are inferring the truth of P in an argument of this form—the "inference" is totally superfluous (Hitchcock, 1992). More generally, since for the realist the truth of the explanans is a requirement on explanation, if we have good reason for believing that the second premise of the argument is true, e.g. by "direct experimental testing", we are already committed to the truth of the explanans and we have no need to infer it.

One problem is that the argument as formulated above is not really inference to the *best* explanation; it's just inference to *an* explanation. An inference to the best explanation is an essentially comparative inference; it says that the *explanans* is the best explanation among some set of possible alternatives. Letting E be this set of alternative explanations, we can reformulate the argument as follows:

$$\begin{array}{l} \quad Q \\ \quad \underline{P\, is\, the\, best\, explanation\, of\, Q\, among\, E} \\ \therefore P \end{array}$$

The way we evaluate whether P is the best explanation of Q among E is by ranking the set of candidate explanations according to a hierarchy of explanatory virtues, such as conservatism, modesty, simplicity, generality, and refutability (Quine and Ullian, 1970; Sinnott-Armstrong and Fogelin, 2015). For the argument to go through, we have to assume that the highest ranking alternative relative to the explanatory virtues makes it more likely that that alternative is true (a controversial assumption). The problem is that, independently of whether we think this assumption holds, if the truth of P is not among those virtues, then the sense of explanation employed in the second premise is not the realist sense, so the realist has to reject this premise: for the realist, whether P is a possible explanation at all requires that it be true. What the realist needs is to make the explanation conditional on the truth of P. But now it looks like we're right back where started since the validity of the argument that makes the truth of the second premise conditional on P depends on having independent reason for believing that P is true, in which case, as before, we no longer need the inference.

A realist construal of explanation appears to create trouble for inference to the best explanation. However, there are at least two important lessons we can draw. First, if one has realist commitments where explanation is concerned, as many philosophers—including mechanists—and scientists do, one does not infer the truth of the *explanans* from explanatory reasoning, but rather assumes it at the outset. Second, this means that, if realism about explanation, at least in part, grounds commonsense scientific judgments about the norms of explanation, evaluating whether a model is explanatory requires independent *reason* for thinking the model is true. Of course, there is no empirical measure that guarantees truth, whether it be empirical adequacy or invariance under intervention (Woodward and Hitchcock, 2003), so whether a model is explanatory on this view is not what I have called a measure-relative feature of that model. Rather, the best one can hope for is that an empirical measure constitutes *evidence* that the model tells us what's really going on. In the specific case of cognitive neuroscience, whether a model is explanatory (in the realist sense) will depend at least in part on whether scientists can generate evidence that a model of cognition tells us something true (or close enough) about what cognition is really like.[13] So part of the reason for thinking that cognitive neuroscience has achieved explanatory success or that it has significant explanatory promise is plausibly based on the judgment that it has successfully identified methods that bear on whether candidate models of cognition tell us just this. Two questions arise: first, what kinds of models are amenable to the evidential reasoning that allows us to sustain realist commitments with respect to those models; and second, what is the

[13] I don't want to say that it has to tell us everything that's going on—indeed, I explicitly reject this as a requirement on explanation in Chapter 4.

nature of the evidence that figures in that reasoning? Consider each question in turn.

For Cartwright, causal models, i.e. mechanistic models, are a class of models that sits comfortably with realism. On her view, one of the reasons causal explanation is so appealing to the realist is that, if we have good reason for thinking that P causes Q, e.g. because we have empirical evidence that this is the case, then we are compelled to believe in the existence of the cause. Thus, on her view, the role of evidence is to establish a causal claim from which we can infer the reality of the cause. Cartwright suggests that this feature, that "existence is an internal characteristic of causation", is peculiar to causal claims. But given that, as I argue below, the actual practice of scientists condones a variety of real properties of systems that figure in our explanatory practices, we ought to consider, if we are going to follow Cartwright's reasoning, that there is a whole family of relations that have this feature, relations that once established, compel us (or if "compel" is too strong, give us reason) to believe in the existence or realness of the relatum. In additional to *cause*, there will be *cause**, *cause***, and so on. And it will be these relations that, for the realist at least, will be contenders for figuring in our best explanatory practices, assuming we can find a source of evidence that these relations actually hold. The point is that there are a variety of real properties of systems that scientists investigate, not all of which are causal, but at least many of which are actual properties of a system that are related in some important way to the observed behavior of that system. And, as in the causal case, we are rational to think this *based on the empirical evidence.* If that is right, then part of the project of developing explanatory models of cognition is going to involve establishing a source of evidence for the truth or accuracy of models that describe a variety of real properties that figure in explanatory judgements, e.g. cause, cause*, cause**. While this is an argument for a kind of explanatory liberalism, it doesn't answer the specific question of *which* kinds of models are amenable to evidence and what that evidence might be. But that is a question for the sciences. And when we turn to cognitive neuroscience, what I think we find is that one powerful source of evidence for a variety of kinds of models is mechanistic evidence. Specifically, by discovering the mechanisms responsible for cognitive phenomena, neuroscientists are able to marshal evidence in support of a variety of kinds of models of those phenomena. The point is not a new one. As Cummins writes:

> Neuroscience enters the picture as a source of evidence, arbitrating among competitors, and ultimately, as the source of an account of the biological realization of psychological systems described functionally. (Cummins, 2000, p. 135)

I set out to say something about how to assess whether a model is explanatory, in particular what grounds commonsense scientific judgments about the norms of explanation. In Chapter 2, I argued mechanism does not adequately account for these judgments across a range of explanatory contexts, but rather accounts for the judgments scientists make when they seek causal-mechanistic explanations. However, a commitment that cuts across commonsense judgments about the norms of explanation in a variety of contexts, one that mechanists appear to share with other philosophers and scientists, is a commitment to realism that is assumed at the outset, rather than as a consequence of explanatory reasoning. On this view, a minimal criterion for the explanatory success of a model is that we have empirical evidence that the model describes actual features of a system that are relevant in some way to the observed behavior of that system. How does the neoclassical model fare on this criterion?

From a descriptive perspective, the predictive success of the model is inconclusive. Although, as discussed above, human behavior deviates from the model's predictions under certain ranges of circumstances, this does not necessarily count against the explanatory success of the model, particularly given the model's record of genuine predictive success under ranges of circumstances from both the individual and aggregate perspectives. But neither is this (albeit limited) empirical adequacy evidence that the model *is* explanatory. The point is that it might be—what we need to show this is empirical evidence that this is the case. The same goes for whether the neoclassical model is an explanatory model of rational choice. In both cases, what we need is evidence that the model describes genuine properties of the system that are relevant to the observed behavior. Where I think cognitive neuroscience has succeeded in finding that evidence is in the neurobiological mechanisms responsible for cognition and behavior.

3.4 Neuroeconomics

The system that guides decision-making in humans and non-human primates consists of two functionally individuated subsystems, a value system that constructs and stores a representation of value, and a "choice network" that recruits this information for action. Consider the value system, which consists of a network of interconnected regions primarily in the basal ganglia of the midbrain and regions of frontal cortex.[14] The value system, as it turns out, is well-described according to a temporal difference (TD) learning model, the goal of which is to successfully predict and store information about future rewards. This system is implemented in the brain in functionally specialized populations of neurons that respond to the neurotransmitter dopamine to generate a reward prediction error (RPE) in response to unexpected rewards, where an RPE represents the difference between the expected value of future rewards and the reward that is actually obtained (or signaled that will be obtained).

A crucial difference between the neoclassical and neuroeconomic models con-

[14] The dopaminergic system is primarily in dorsal and ventral regions of the ventral tegmental area (VTA), substantia nigra pars compacta (SNc).

cerns the reference-dependence of the value signal encoded in an RPE. Neoclassical economic theory requires that a rational decision-maker be able to represent the *objective* value of rewards in a choice set and then compute the marginal utility of those rewards from the choosers utility function and the state of the chooser's wealth (Glimcher, 2011). But animals are (almost certainly) incapable of representing an objective measure of value.[15] Rather, they construct a representation of value relative to a reference point that is constantly shifting depending on the animal's circumstances. This inevitably leads the brain to make mistakes in assessing the value of goods under certain conditions. However, encoding value relative to a reference point is an optimizing encoding strategy under constraint, namely the high (i.e. disproportionate) metabolic cost of neurons.[16] Given these metabolic constraints, by encoding representations of value relative to a baseline, the populations of neurons that encode these representations are able to represent a far greater range of feature magnitudes than they otherwise would be able to do.[17] On the other hand, this can lead to errors in judgment in circumstances where there are significant shifts in the reference point.

Dopamine neurons project widely throughout the brain, particularly to regions of the choice network in frontal and parietal cortex that are responsible for generating musculoskeletal and eye movements. The value signal originating in the value system

[15] One source of evidence for this conclusion is from decades of research on the sensory modalities. Animals don't represent an objective measure of perceptual features, e.g. luminosity, acoustic amplitude, salinity or sweetness, somatosensory properties (pain, pressure, temperature), etc. Rather, sensory neurons only encode a measure relative to a reference point, and they do so as a compressive power function of the actual feature magnitude. Since the sources of the brain's representation of value are the sensory regions, it makes sense that this representation is constructed under the same constraint.

[16] As Glimcher notes, the brain consumes around 20% of an adult human's metabolic resources, despite only accounting for 4% of the weight of the human body. Since it is expensive to increase the number of neurons, an alternative solution is to increase, or even optimize, the information storage capacity of the neurons you do have.

[17] This reference-based measure of value has been modeled in economic theory as relative expected value. (McFadden, 2005)

is encoded in this choice network in topographically organized populations of neurons, or neural "maps".[18] As an example, consider the projections from the retina to topographically organized maps in the lateral geniculate nucleus (LGN): "A pattern of action potentials at a single location on this map reveals not only the location of photons in the visual world but also wavelength distribution of those photons and information about the spatial and temporal frequencies of the patterns of light and dark in the outside world." (Glimcher, 2011) The process whereby this encoding occurs is a remarkably elegant design solution to the problem of assigning expected values to actions associated with options in a choice set. Dopamine is known to be a key component in the strengthening of synaptic connections between the axons and dendrites of cortical neurons.[19] Whenever a movement is rewarded in a way that differs from expectations, an RPE is generated by the value system and the neural maps in the choice network are bathed in dopamine, which strengthens the synaptic connections at the location in the map that corresponds with the rewarded movement. This process thus encodes the expected value of an outcome in the mean firing rate of cortical neurons associated with the movement that generates the valued outcome. In this way, the expected value of an outcome in a choice set gets paired with the motion-behavior that will achieve that particular outcome. There are two important features of how this information is encoded in the choice network that have important ramifications for the choice behavior we observe.

First, the magnitude of the encoded expected value is normalized relative to other options in the choice set. This allows the choice network to make precise discrimina-

[18] These maps are found all over the place in the brain. They constitute a high-dimensional information encoding mechanism whereby a feature magnitude is associated with a location on a neural map that corresponds with, for example a target location in visual space, a location on the body, or a particular eye or musculoskeletal movement. This topographic organization of information is, as Glimcher writes, "a basic organizing principle of the brain" (Glimcher, 2011, p. 147).

[19] Known as long term potentiation (LTP).

tions in the value of options in the choice set, but at the cost of limiting the number of options that can be effectively discriminated. As with the reference dependence of the value signal, this trade-off between discriminatory power and encoding capacity is an optimizing encoding strategy given the metabolic (and physical) constraints that our evolutionary ancestors encountered. However, this limitation leads to errors. Think of a neural topographic map as a noisy information channel with limited capacity for encoding the expected value of options in a choice set. Since the channel is limited, as the number of options in the choice set increases, the noise in the signal will eventually prevent a discrimination in firing rates. And the "noise" or stochasticity of cortical neurons is very high. Specifically, the stochasticity of the interval between synapses is precisely characterized by a modified Poisson distribution function which has a variance very nearly equal to the mean.[20] This stochasticity places a significant constraint on the number of options in a choice set that can be effectively represented. However, rather than constituting a mechanistic limitation of neurons, and this is a second important feature of how expected value is encoded at various stages in the choice network, this stochasticity serves an important computational function, and represents another design trade-off in solving the various decision problems that an animal might face. If the firing rate of adjacent neurons in a population that encodes the expected value of an option is effectively uncorrelated, then the high stochasticity of their firing rates serves as a vector averaging mechanism that is able to extract with very high precision the mean firing rate of the expected value signal. However, the really remarkable thing is that this interneuronal correlation can be modulated depending on the decision-problem encountered. When the firing rate of adjacent neurons in a population is correlated, rather than encoding a precise representation of the mean they will instead encode the stochasticity of the afferent signal so that the expected values represented in the final stages of the choice net-

[20] It is modified due to a mechanistic limitation on the rate at which neurons can synapse.

work fluctuate randomly, according to the stochasticity of the afferent signal. There is evidence to suggest that encoding stochasticity into the signal that guides behavior corresponds to an optimal decision strategy in certain contexts, particularly in contexts that involve the kind of social decisions modeled in game theory.

Of course, one is always going to have limitations on the representational capacity of a noisy information channel so that eventually we should expect to see errors when this representational capacity is reached. However, these potential errors are mitigated—in decision contexts that our evolutionary ancestors encountered—by the optimizing encoding strategies that were selectively advantageous, strategies which, for example, value discriminatory power over breadth of options, short over long-term rewards, etc. We expect to see these limitations and the encoding strategies that the brain adopts to mitigate these limitations in the choice behavior and this is exactly what we do see in the experiments described above that show deviations from the neoclassical model's predictions.

The final stage of the decision process involves generating the action that has the highest expected value. As the choice network receives additional information about the likelihood of future rewards from the dopaminergic system, and the expected value encoded in the mean, normalized firing rate of cortical neurons increases for a given option in a choice set, inhibitory neural connections to distant locations in the neural map suppress activity at those locations. In this way, outcomes in a choice set compete for control of the regions of the brain responsible for generating movement. When a biophysical threshold is reached in the mean firing rate at a location in the cortical map, neural activity in the regions responsible for control transitions irrevocably to a bursting mode that generates the motion associated with the highest expected value.

3.5 Historical development of the model

Researchers investigating the dopaminergic system had known for some time that dopamine neurons responded in their firing rates to the receipt of rewards. However, there was little agreement as to what functional role these neurons played. Mainly, researchers thought dopamine somehow responded to pleasurable experiences. In the early 1990s, Schulz and colleagues performed an important series of experiments on monkeys, based on the well-known conditioning experiments pioneered by Pavlov, in which they recorded the activity of dopamine neurons as the monkeys were conditioned to expect a reward after hearing an unpredictable tone.

To illustrate, consider a simple scenario. Call it the "100% reward condition". A monkey hears an unpredictable tone at T_0. Five seconds later at T_1, the monkey receives a juice reward. This is repeated several times—an unpredictable tone at T_0 followed every time by a juice reward five seconds later. Shultz and colleagues observed that, initially, dopamine neurons responded with an increase in firing rate at T_1. However, with successive trials, the time at which the dopamine neurons fired propagated backwards from T_1 to T_0, so that eventually the dopamine neurons increased their firing rate only in response to the tone, and not the juice reward. Montague, then working with Shultz, recognized this pattern of behavior as being predicted by the TD learning model, which Sutton and Bartow had recently developed as an optimal machine learning strategy. The goal of a TD learning system is to predict the value of future reward by updating the current expectation of reward based on new information. Specifically, Montague proposed that dopamine neurons encode a reward prediction error, which in the TD model, is the difference between the expected value of future rewards and the reward that is actually obtained. In this way, dopamine neurons, which project widely throughout the brain, project information about expected value to a whole suite of systems, including those responsible for generating musculoskeletal and eye movements.

Investigation of the choice network was largely independent of research on dopamine. A significant breakthrough came with Shadlen and Newsome's investigation of areas responsible for eye movements (areas MT and LIP) (Shadlen and Newsome, 1996). In particular, they developed a method for testing the hypothesis that neurons in this region predict eye movement. In this random-dot discrimination task, a monkey views a display with small dots randomly moving around the screen, but with some fixed percentage of the dots exhibiting leftward or rightward motion. The monkey's task is to decide which direction the dots are moving by saccading to a target in the direction of perceived motion. If the monkey succeeds then he receives a juice reward. Shadlen and Newsome recorded the activity of neurons in area LIP as the monkey performed this task and discovered that activity in this region precisely predicted the animal's choice behavior. Specifically, as the monkey gained more information from the random-dot display about the direction of motion, neurons in LIP associated with the rewarded movement increased their firing rates over time. They interpreted these data from a stimulus-response perspective as showing that neurons in area LIP associated with a particular stimulus-response pair compete for control of the saccadic control system.

However, in their 1999 paper, Platt and Glimcher interpret these data as showing that expected value is being encoded in area LIP and that what better *explains* the movement than a stimulus-response description of the mechanisms in the choice network is that the monkey moves his eyes in the direction that has the highest expected value associated with that movement. Indeed, this explains not only the behavior of the monkey but also why the monkey's brain consists in these mechanisms: because tracking the expected value of goods is an effective strategy for maximizing future rewards. The significance of this proposal must not be underrated. It unites several seemingly independent research programs across multiple disciplines into a single research agenda, organized around a particular model that has at its core the idea that the brain guides choice by constructing a representation of expected value for various kinds of goods or outcomes and then generating the behavior that has the highest expected value of the outcomes in a choice set. The upshot is that the notion of expected value, a core concept in neoclassical economic models of decision-making, takes on paramount *explanatory* significance with respect to animal decision behavior.

In Chapter 2, I described two accounts of mechanism, a strong metaphysical thesis, and a weaker methodological thesis. According to the former, mechanisms are systems composed of entities that perform activities, where activities are the causal dependencies between entities that are responsible for the occurrence of the explanandum. On this view, an explanatory model or feature of a model is not mechanistic if it meets one of the following conditions: the feature is not included in a description of the entities that compose the mechanism and the causal relations among those entities; or if the feature can be characterized independently of the mechanism(s) that instantiates it. According to the methodological thesis, mechanistic explanations result from a conceptual decomposition of a system into parts that perform operations. That is, mechanism amounts to an account of a particular kind of explanatory practice. On this view, a model is not mechanistic if it is not the result of this kind of practice. In the remainder of this Chapter, I assess each of these theses against the description of the neuroeconomic model and its historical development described above.

3.6 Assessing strong mechanism against the neuroeconomic model and its historical development

Cognitive neuroscience engages in mechanistic explanation of the kind that Craver and Kaplan endorse. Consideration of the neuroeconomic model shows this. Mechanistic forms of explanation appear to be especially prevalent in this model in explaining how the choice network recruits economic information for action—for example, by describing how the RPEs generated by the value system produce a bath of dopamine in the topographic maps in the choice network, strengthening the synaptic connections at locations in the map that correspond to the rewarded movement. I take it that this is an archetypal mechanistic explanation. But this is neither a surprising nor controversial point either from the perspective of the neuroeconomic model or the philosophy of explanation. Our explanatory goals will include knowledge of the mechanisms responsible for encoding value and recruiting that information for action, the implementation details of how neurons form circuits, the nature of the neural circuits that encode information, how they do this, and their limitations. From a philosophical perspective, given that a physicalist metaphysics endorses causal-mechanistic relations as one of the principal relations in a physicalist ontology, we should expect that our explanatory knowledge of the world, to include investigation into the nature of cognition, will include knowledge of those relations. Both of these limited claims, I take it, are uncontroversial. However, in other important respects, the neuroeconomic model is poorly characterized by mechanist commitments. In this section I consider one important respect in which this is true. I then also consider how the historical development of the neuroeconomic model supports this claim.

3.6.1 Neuroeconomics as an information-processing model

A science of the mind is, at least in part, concerned with investigating how cognitive states are related to the distal environment in such a way that a cognitive agent is able to function in the agent's environment (Bechtel 2008, p. 27). One way to characterize this relationship is by hypothesizing that an animal's cognitive system functions by extracting information from sensory inputs to construct *representations* of salient features of the environment, and then "transforming" these representations to guide the animal's behavior. On this information-process approach to explaining cognitive capacities and behavior, one explains a cognitive capacity or behavior by citing the representational features of cognitive states that are relevant to explaining how those states guide behavior. To more precisely characterize these explanatorily relevant features, we need a more specific characterization of representational states. Specifically, these states of a cognitive system can be characterized as having a bearer and a content. The bearer is a physical state of the system that instantiates the representation. The content is that which the representational state represents, for example, some feature of the system's distal environment. Given this construal of representations, we can characterize information-processing as successive transformations of the information bearer in such a way that the transformations correspond to meaningful operations (for example, logical or mathematical operations) on the contents of the bearers.[21] An information-processing model thus *explains* behavior by citing the content of information-bearing states and the operations the system performs on those states such that the system is able to guide behavior appropriate to the system's environment.

Consider, as an example, how an information-processing model explains a simple reaching movement to grasp a familiar object in visual space. How does the cognitive system accomplish this? On an information-processing model, by extracting information from a two-dimensional stereoscopic retinal image, the brain constructs representational states about the object, its features, and location, and exploits these representations to generate the appropriate reaching movement. For example, the brain must compute a trajectory for the hand, which is at a different location relative to the object than the eyes; as well as appropriate grip strength and hand rotation

[21] There is an ambiguity in the term "computation". We can understand computation as a purely syntactic operation over strings of symbols (changes in the information-bearer), or as the interpretations that we assign to those operations, for example, as performing a particular mathematical or logical function.

for objects of this shape, size, and density. This is accomplished via transformations in the physical states that bear the information about the object (synaptic activity in populations of neurons) that correspond to meaningful operations on their content, for example, a coordinate transformation function, instructions for grip strength, etc. In this way, a person's cognitive states are related to the external environment in such a way that the person can successfully reach for and grasp the object.

Two crucial features of an information-processing model are meaningful cognitive states (states that have content), and the meaningful computations or transformations performed on those states. Thus to fully characterize an information-processing explanation, we would need the following:

- A theory of meaning: how an information-bearer can come to have content;
- A theory of computation: how changes in physical states can correspond to meaningful transformations in, or operations on, contents;
- A theory of explanation that tells us how meanings and computations are relevant to explaining the behavior of the system.

Fortunately, for my purposes here, I do not need to address any of these issues directly. Rather, it will be sufficient to show that mechanism, as a theory of explanation, does not adequately account for what is required in a plausible theory of meaning and computation.[22] The argument proceeds as follows:

1. According to strong mechanism, what is relevant to explaining the behavior of a cognitive system are the nested causal dependencies that jointly produce the explanandum, where causal dependencies are intrinsic properties of the cognitive system.

[22] Van Eckhardt and Poland make this point as well (Von Eckardt and Poland, 2004).

2. On an information-processing model, what is relevant to explaining behavior are the contents of information-bearing states and the operations that are performed on those states to guide behavior, i.e. what is explanatorily relevant are the semantic properties of the information processing system.

3. But according to a widely help view about meanings, semantic properties are not intrinsic properties of the system—they are relations between physical states and features of a cognitive agent's distal environment. So, those features of an information-processing model that are relevant to explaining an animal's behavior are not included in a mechanistic model.

4. Therefore, information-processing models are not mechanistic models.

The neuroeconomic model of decision-making is an information-processing model *par excellence*: it describes how the brain constructs and stores a representation of the value of various outcomes, and how this information is transformed in such a way as to guide behavior. More specifically, it describes how the value system constructs a representation of the values of a set of feasible alternatives, encoded in the mean firing-rate of populations of neurons. By generating an RPE, the brain then encodes the relative expected value of a set of outcomes into the choice network, pairing these with motor responses that will achieve the associated outcome. But is the model explanatory?

There are really two questions here. The first is whether information-processing models generally are explanatory. The second is whether this particular information-processing model is explanatory of human choice behavior. If the answer to the first question is yes—and I do not find persuasive mechanist arguments that they are not, then it looks like the answer to the second question is also yes. In Section 3 above, I outlined a minimal criterion for explanatory success that expands on mechanism's own realist commitments regarding the nature of explanation: that we have empirical evidence that the model describes actual features of a system that are relevant in some way to the observed behavior of the system. Cognitive neuroscience has been able to marshal mechanistic evidence in favor of a neuroeconomic model of human choice behavior. For example, the Shadlen-Newsome random dot discrimination task is evidence that the brain exploits representations of the expected value of outcomes to make choices. It was with this work as a foundation that Platt and Glimcher proposed that these mechanisms were suitable for implementing a system that recruits information about expected value to guide saccades. That is, the discovery of these mechanisms in the occulomotor choice network were both a crucial component in the development of the neuroeconomic model, and in marshaling evidence in its favor.[23] Another source of evidence is Shultz's work on single-unit recordings in dopaminergic pathways during conditioning tasks. This work specifies mechanisms that are evidence that the brain constructs representations of value and broadcasts this information to parts of the brain responsible for choice behavior via the propagation of a reward prediction error.[24]

[23] This is not to say that predictive success cannot be evidence of explanatory success. For example, predictive success is powerful evidence of explanatory success in physics. The science of mind is complicated however by the sheer complexity of the system under study and the fact that no single system is exactly alike. So, for example, given these complications, it is difficult in a science like cognitive neuroscience to assess the significance of predictive anomalies. Not so in physics where measurements to an incredibly high degree of precision are both possible and routinely employed in evidential reasoning.

[24] I have been discussing research on how monkeys move their eyes in response to juice rewards and I have said that explaining how the monkey does this and why it does it in one way rather than another is that the monkey's brain constructs a representation of value and then exploits this information to guide choice. One philosophical worry is that this research has nothing to do with *real* value or *real* choice, and so has nothing to do with *real* decision-making. That is, even supposing the neuroeconomic model does explain how monkeys move their eyes, it does not explain the kind of decision-making that we are interested in when we think of the kind of *mental* behavior that philosophers are actually interested in explaining. There are two replies. The first is that we have good reason to believe that this model will scale up to "real" decision-making. One reason is precisely because this research employs an intentional notion. In this respect, philosophers who have bemoaned appeal to the neurosciences to solve philosophical problems are partially vindicated. In fact, neuroscience has adopted philosophical, i.e. economic notions in order to explain how the brain guides decisions. The second reply is to point out that my target here is the mechanist thesis

Another important way in which information-processing models explain is in how certain information-processing strategies can be characterized as effective or optimizing strategies given, for example, the biological constraints of the information-processing system (the brain). These correspond closely with what Collin Rice has called optimality explanations, according to which the "goal is to identify which values of some control variable(s) will optimize the value of some design variable(s) in light of some design constraints" (Rice, 2013, p. 3). On the neuroeconomic model, the brain adopts certain information-theoretic strategies that mitigate limitations imposed by evolutionary and biophysical constraints, for example, broadcasting a reward prediction error to encode expected value; encoding expected value relative to a reference point to mitigate metabolic constraints; and normalizing the value of options in a choice set to increase discriminatory power in a noisy information channel.[25] In fact, remarkably, Sutton has shown that the TD learning model is a provably optimal learning strategy compared to alternatives (it converges to the correct answer faster and with fewer steps) (Sutton 1988). That the brain adopts these strategies explains why we observe certain kinds of behaviors. For example, we can explain why subjects choose randomly when the size of a choice set exceeds approximately eight options. This occurs because the brain normalizes the expected values of options in the choice set, which increases discriminatory power when the

described in Chapter 2. Even supposing that the neuroeconomic model doesn't explain the kind of thing philosophers are interested in explaining, it remains the case that the neuroeconomic model explains a type of behavior that is a core concern of cognitive neuroscience and it does so in a way that is inconsistent with mechanism. In other words, whether or not one thinks this argument has significance for the philosophy of mind, it does for the philosophy of science: a core area of cognitive neuroscience achieves explanatory success and it does so non-mechanistically. That said, I concede that it remains an open possibility that cognitive neuroscience, at least as far as real decision-making is concerned, simply does not (yet) have the resources to answer the questions philosophers are interested in answering. However, I will not address this concern here.

[25] The encoding of expected value in the mean firing rate of highly stochastic populations of neurons points to another example of non-mechanistic features of the neuroeconomic model, namely, how statistical models of the stochastic properties of the populations of cortical neurons in the choice network are central to understanding certain kinds of choice behavior. However, I leave further development of this example for future consideration.

number of choices is low, but produces errors as set size increases depending on the capacity of the information channel (which for a firing neuron is limited by a synapse frequency range of approximately 10 to 100 Hz). Furthermore, these strategies can be characterized independently of the mechanisms that realize them. In fact, understanding these information-theoretic strategies explains why one finds the mechanisms one does in the decision-making system. Normalization is an optimizing strategy for mitigating the limitations on the information capacity of the channel relative to the environment in which these systems evolved, an environment in which, for example, greater discriminatory power was more selectively advantageous than the ability to consistently choose from a large set of options.

Paul Churchland, following Sellars, characterized belief-desire psychology as a folk theory that we use to predict and explain human behavior. He has argued that, given its empirical failures, this theory is actually a pretty bad theory, on a par with other abandoned scientific theories. In response, he bet that a future neuroscience would dispense with intentional notions altogether in favor of, for example, a language of synaptic activity, feed-forward neural networks, and vector-prototype activations (Churchland, 1989). But in fact, at least in the neuroscience of decision-making, this bet has not been borne out: our best science of human choice behavior remains shot through with intentional notions. Furthermore, we now have a much better understanding of why people make decisions that conflict with economic theory. Indeed, the neuroeconomic model gives us a theoretical framework for assessing the rationality of actual human choice behavior.

3.6.2 Historical considerations

The history of how explanatory success came about in the science of decision-making is inconsistent with the historical picture to which strong mechanism is committed, namely, that explanatory progress *consists in* the steady increase in our knowledge of the mechanisms responsible for cognitive phenomena. If Craver and Kaplan are right then we should expect to see progress occur because of mechanistic advancement. But in this case, it was knowledge of the mechanisms that came first, and then the explanations. Specifically, Shadlen and Newsome had developed the basis of a detailed mechanistic model of the occulomotor system. We already had a detailed understanding of how the brain represents discreet dimensions of information in topographically arranged neural maps. We knew the effect of dopamine on synaptic connections and the details of LTP. Significant explanatory progress required a different kind of insight. First, Montague construed the dopaminergic system as realizing a temporal difference learning system, which provided a "computationally tractable" method of encoding value (Smith and Huettel, 2010). Second, Platt and Glimcher hypothesized that the choice network encodes expected value. These hypotheses form the core of the neuroeconomic model—it is only when these insights come together in this way that a leap in explanatory progress occurs. With the neuroeconomic model, we get a framework that provides answers to a whole suite of explanation-seeking question that prior mechanistic models did not, by themselves, have the resources to answer. Furthermore, we get a model according to which these mechanistic discoveries are themselves explained. At the same time, the mechanisms implicated in realizing this system are a powerful source of evidence that the model is a good one, and that it really does explain.

3.7 Assessing methodological mechanism

Does the neuroeconomic model vindicate Bechtel's naturalistic mechanism? The kinds of explanatory strategies that ones finds in the neuroeconomic model constitute perspectives that Bechtel countenances as kinds of mechanistic explanation. As noted in Chapter 2, Bechtel explicitly construes information-processing models as mechanistic models of cognition, so on Bechtel's view the neuroeconomic model is a mechanistic model. What that means for Bechtel is that it is the product of an explanatory practice that proceeds by conceptually decomposing a system into parts that perform operations. There are two points I wish to make.

The first is that, on Bechtel's view, mechanistic explanations are representations of a system (mechanistically understood) from different perspectives or "stances". But on that view what does the bulk of the explanatory work is not *that* they are descriptions of such systems, but rather the kind of description they are, for example, information-processing descriptions, statistical descriptions, or mechanistic descriptions (in the sense of Craver and Kaplan) of the system in question. If that is right, then it looks like what we have uncovered, through a methodological commitment to mechanistic decomposition, is actually different kinds of explanations.

Second, even supposing it is right to think that the brain's being amenable to mechanistic decomposition has explanatory significance, this is not the only way of thinking about the brain. Sometimes the brain is like the weather, and the fact that scientists treat it that way when they use the tools of dynamical systems analysis to understand how the brain works suggests that perspectives on the brain in cognitive neuroscience that lead to our understanding of it are not limited to the mechanistic perspective. Given Bechtel's naturalism, this means that not only should Bechtel admit different types of explanations discovered via mechanistic decomposition, but that he also ought to concede that cognitive neuroscience adopts explanatory practices other than the mechanistic account he defends.

That said, I think there is an important insight in Bechtel's methodological approach. If we interpret this approach as a methodological commitment to the evidential role of mechanisms, in something like the way behaviorism consisted in a methodological commitment to the evidential value of behavioral responses to stimuli, rather than as a distinctive theory of how to explain cognition, then his approach takes on a different significance. This is that the discovery and description of mechanisms has had a significant role in the explanatory success of cognitive neuroscience, not because mechanistic description is what explanation of cognition necessarily consists in, but rather because it is a rich source of evidence in theorizing about cognition.

3.8 Conclusion

One can imagine that the science of mind would have pursued an ever more reductionist investigatory strategy in the pursuit of explanatory success. But this has not been the case in research on the neuroscience of decision-making. Instead, somewhat surprisingly, cognitive neuroscientists have co-opted concepts from such "higher-level" disciplines as economics, cognitive science, engineering, and psychology in order to explain what's going on in the brain. This strategy, I have argued, is inconsistent with the norms embodied in the strong mechanist program that Craver and Kaplan defend. In virtue of what then are the models of cognition in cognitive neuroscience actually explanatory? In Chapter 4, I approach this question by considering the relationship between explanation and understanding.

4

Objectivism

4.1 Introduction

Cognitive neuroscience explains by constructing a variety of kinds of models. While some of these models are "mechanistic", not all of them are, nor are they primarily so. Instead, the principal role of mechanistic discovery is evidential: mechanisms are marshaled as evidence in favor of explanatory models of cognition. In virtue of what then are these models explanatory? What, if anything, unifies these various kinds of models as instances of genuine explanation?

In his essay, "The Importance of Scientific Understanding," Wesley Salmon notes that despite several decades of attention to (and progress in) developing accounts of scientific explanation, philosophers have paid little attention to what Carnap called "clarification of the explicandum" (Carnap, 1974), which is to say, attention to "the value of scientific explanations or of the reasons for seeking them." (Salmon, 1998, p. 82) To address this shortcoming, Salmon (in this paper, as elsewhere (Salmon, 1984)) invokes what he takes to be the "key concept" required to elucidate the value of explanation: explanations are worth seeking because they produce *understand-*

ing.[1] His stated aim in this essay is thus to "characterize the kind of intellectual understanding we can achieve."

Let us suppose that the value of explanatory models in cognitive neuroscience is to be found, as Salmon claims, in the scientific understanding they produce. How then to characterize understanding? In this Chapter I evaluate an approach to understanding that has dominated the philosophy of science since Hempel's seminal work on scientific explanation, one that reaches its fullest formulation in Salmon's causal-mechanistic theory. According to this view, henceforth "objectivism", understanding is a traditional form of knowledge, namely, knowledge of certain "objective" relations between the phenomenon to be explained (the *explanandum*) and what does the explaining (the *explanans*).

This chapter has two parts. In the first part, I trace objectivism to Hempel's formulation of the covering-law account of explanation and show how Salmon extends this basic approach to the causal-mechanistic accounts of scientific understanding that he defends. I discuss understanding in the specific case of the mind sciences by considering a modern day formulation of objectivism in the work of Carl Craver that is aimed specifically at understanding in the cognitive neurosciences. In the second part of this chapter, I argue that objectivism does not adequately account for the ends we actually value or the ends we ought to value when we seek understanding.

4.2 An Example

Before continuing, it will be helpful to have an example that illustrates the basic idea behind objectivism. This is the Monty Hall problem. Consider a fictional version

[1] Salmon writes: "Perhaps the most important fruit of modern science is the understanding it provides of the world in which we live, and of the phenomena that transpire within it. Such understanding results from our ability to fashion scientific explanations...I raised a fundamental philosophical query about the nature of scientific explanation, namely, what sort of knowledge is explanatory knowledge, and on what basis can we say that it constitutes or contributes to understanding?", then adding that this question "has received surprisingly little explicit attention in the literature" (Salmon, 1984, 259).

of the gameshow Let's Make a Deal, in which Monty Hall gives a contestant an opportunity to choose between three doors, two of which conceal a goat, and one that conceals a brand new sports car. The contestant is directed to choose one of the doors without opening it. The contestant chooses, at random, one of the doors. Mr. Hall then opens one of the unchosen doors that has a goat behind it and gives the contestant a choice: she can either stick with her original choice, or switch to the remaining unopened door. What should the contestant do to give herself the best chance of winning the car?

Most people think that it does not matter which door the contestant chooses. In fact, most people's intuitions about the right answer to this problem are wrong. The contestant should always switch doors to maximize her chances of winning the car. This counterintuitive fact cries out for an explanation. What explains why the contestant should switch doors to maximize her chances of winning the car? Since the question concerns the probability that a car is behind one of two doors, *given that* Mr. Hall revealed a goat behind a third door, this is a conditional probability problem, and so the probability of switching versus not switching can be calculated with the appropriate formula for calculating conditional probabilities, namely Bayes's theorem, a theorem of the axioms of the probability calculus. The theorem tells us the probability of A given B as follows:

$$P(A \mid B) = \frac{P\,(A)\,P\,(B \mid A)}{P\,(B)}$$

Applied to the Monty Hall problem, we get the following. Given the set S={1,2,3} each member of which represents one of the three doors, and given that:

C is the door that conceals the car

X is the door that the contestant chooses

M is the door that Monty opens

then the probability that C is the door that neither the contestant nor Monty chose (the door potentially switched to) can be calculated as:

$$P(C = 2|M = 3, X = 1) = \frac{P(M = 3, C = 2|X = 1)P(C = 2|X = 1)}{P(M = 3|X = 1)}$$

$$= \frac{P(M = 3|C = 2, X = 1)P(C = 2|X = 1)}{\sum_{i=1}^{3} P(M = 3|C = i, X = 1)P(C = i|X = 1)} = \frac{2}{3}$$

while the probability that $C \neq 2$ given that $M = 3$ and $X = 1$ is just $1 - P(C = 2|M = 3, X = 1)$, or $\frac{1}{3}$. Since the probability that the contestant will win the car if she switches is $\frac{2}{3}$, but only $\frac{1}{3}$ if she doesn't, the contestant actually doubles her chances of winning the car if she switches doors.[2]

According to objectivism, if this explanation is both true and complete, then the having of it amounts to having genuine "scientific" understanding of the problem, regardless of whether any particular individual comprehends the solution, or finds the solution intelligible. This is the answer to the problem, and the explanation provided explains why the answer is what it is, regardless of whether one finds the answer intelligible or not. And in this case, most people do not find the answer intelligible—in fact what people do find intelligible is exactly the wrong answer to the problem.

To make matters worse for any psychological notion of understanding, making this problem intelligible requires describing it in different ways—that is, what's interesting about this problem is that what "explains" the problem (in the psychological sense) appears to vary from individual to individual. Here is one explanation:

[2] Source: http://en.wikipedia.org/wiki/Monty_Hall_problem, 22 Feb 2015

> Consider the only three possible scenarios that could occur: either the contestant first chooses the door that conceals the car, or the contestant chooses the door that conceals goat #1, or the contestant chooses the door that conceals goat #2. In each scenario, Monty Hall opens one of the remaining doors that conceals a goat. Consider what happens in each scenario if the contestant switches doors. In the first scenario, she switches from the car to the remaining goat. But in both the second and third scenarios, she switches to the door concealing the car. So in two out of three possible scenarios, the contestant wins the car when she switches, compared to only one of three when she doesn't, namely when she initially chooses the car. So, the contestant doubles her chances of choosing the car when she switches doors.

Here is another explanation:

> Imagine that instead of only 3 doors, there are 1000 doors, only one of which conceals a car. The contestant chooses one door at random. Monty Hall then proceeds to open 998 doors revealing 998 goats. How likely is it that the contestant initially chose the car? Now do you think the contestant should switch doors?

Objectivism about understanding is, on the face of it, an appealing approach to how to think about explanation and understanding in this case. And to the extent that this example bears a structural resemblance to other problems in the sciences—quantum mechanics and the problem of consciousness come to mind, it makes objectivism an appealing approach more generally. I turn to a discussion of the historical origins of this view and how it has become the predominant view of scientific understanding.

4.3 Objectivism about understanding

It is a curious fact that a second wave of positivist-oriented philosophers (Hempel in particular) concerned themselves with articulating a philosophical model of explanation (Woodward, 2011). The basic tenets of logical empiricism seem to exclude just such a notion, which is, historically speaking, either overtly metaphysical insofar as it has an air of attempting to get behind the appearances to the essence of things, or overtly psychologistic in its appeals to psychological notions like "intelligibility" and "understanding"—the very things that positivism was at pains to overcome in developing a respectable epistemology. Duhem's explicit denial of any "explanatory" role for physics is characteristic of early positivist skepticism towards the metaphysical commitments of explanation (Duhem, 1991). And Frege's objectivist response to Kantian psychologism and its heirs, particularly Mill's quasi-psychologism about logic (Mill, 1956), provide a model for logical empiricist anti-psychologism in epistemology generally, and the epistemology of science especially (Sober, 1978). But an alternative rhetorical strategy to insistence on the purely systemizing role of scientific theory, and skepticism concerning the epistemological significance of psychological notions like understanding, is to reconceptualize the nature of explanation along logical empiricist lines, and with it our conception of scientific understanding.

4.3.1 Hempel and scientific understanding

A fruitful way to interpret the Hempelian program is precisely as an effort to make explanation an epistemologically respectable notion.[3] Hempel all but tells us as much when he concludes *Aspects* with the following observation:

> The central theme of this essay has been, briefly, that all scientific explanation involves, explicitly or by implication, a subsumption of its subject

[3] As many have noted, Hempel did not invent deductive nomological explanation. The tradition has roots in Aristotle.

> matter under general regularities; that it seeks to provide a systematic understanding of empirical phenomena by showing that they fit into a nomic nexus. This construal...does not claim simply to be descriptive of the explanations actually offered in empirical science; for—to mention but one reason—there is no sufficiently clear generally accepted understanding as to what counts as a scientific explanation. The construal here set forth is, rather, in the nature of an *explication*, which is intended to replace a familiar but vague and ambiguous notion by a more precisely characterized and systematically fruitful and illuminating one. (Hempel, 1965, 489)

Physics, insofar as it appeals to laws, is ideally suited for such a project. Laws, after all, are a respectable empiricist device when properly conceived: they are systematizing descriptions of the appearances, so there is no question of their falsely representing the *true* nature of reality—in the logical empiricist sense, they make no claims to. Indeed, one of the hallmarks of Newton's transformation of the scientific enterprise was his positing of forces, and the laws that govern them, that saved the appearances without "feigning hypotheses" concerning the metaphysical status of either (Newton, 1999).[4] If the laws correctly systematize, then they perform their proper epistemic function, one that is consistent with the more extreme versions of positivist instrumentalism: there might be multiple but potentially inconsistent systematizations that nevertheless do the job.

One very important feature of Hempel's account of scientific explanation is its anti-psychologism. Hempel is explicit in his rejection of any appeals to psychology as relevant to the study of the *logic* of explanation, insisting that psychological features

[4] Though it should be noted that his not feigning hypotheses about the metaphysical status of forces is consistent with interpreting Newton as taking the gravitational force law to be explanatory.

are merely pragmatic aspects of explanation.[5] In a response to Scriven, who argues that a complete answer to an explanatory question ”is one that relates the object of inquiry to the realm of understanding in some comprehensible and appropriate way” (Scriven, 1962), Hempel replies in a well-known passage that ”such expressions as ’realm of understanding’ and ’comprehensible’ do not belong to the vocabulary of logic, for they refer to psychological or pragmatic aspects of explanation.” (Hempel, 1965, p. 413) To illustrate the point, he provides an instructive comparison to the notion of mathematical proof:

> The case of scientific explanation is similar [to the case of mathematical proof]. For scientific research seeks to account for empirical phenomena by means of laws and theories which are objective in the sense that their empirical implications and their evidential support are independent of what particular individuals happen to test to apply them; and the explanations, as well as the predictions, based upon such laws and theories are meant to be objective in an analogous sense. This ideal intent suggests the problem of constructing a nonpragmatic concept of scientific explanation—a concept which is abstracted, as it were, from the pragmatic one, and which does not require relativization with respect to questioning individuals any more than does the concept of mathematical proof. It is this nonpragmatic conception of explanation which the covering-law models are meant to explicate. (Hempel, 1965, p. 426)

[5] Hempel writes:

> In history as anywhere else in empirical science, the explanation of a phenomenon consists in subsuming it under general empirical laws; and the criterion of its soundness is not whether it appeals to our imagination, whether it is presented in terms of suggestive analogies or is otherwise made to appear plausible—all this may occur in pseudo-explanations as well—but exclusively whether it rests on empirically well confirmed assumptions concerning initial conditions and general laws. (Hempel, 1965, p. 240)

Hempel's reference to mathematical proof echoes Frege, who was concerned to show that which mathematical inferences are licensed does not depend on the psychologies of individual mathematicians.[6] Hempel extends this idea to explanation: if explanations have epistemological merit, they must be construed "objectively", which for Hempel, as for Frege, means independently of the psychological characteristics of the individuals who employ them. So whatever explanation is, it must derive its epistemological merit from a legitimate source of objectivity, which in the logical empiricist tradition are limited. In particular, they are limited to statements of empirical observations and the logical relations between them—hence the concern with providing an analysis of the logic of scientific explanation while relegating psychology to its merely pragmatic aspects. For, if Frege is right, then what inferences are licensed from premises does not depend on the particular individuals who make those inferences, but rather is a matter of pure logical necessity. Thus, according to the "covering-law" model that Hempel defends, explanations are *arguments*, in which the *explanandum* phenomenon is a licensed inference (deductive consequence in the case of the deductive-nomological explanation, and inductive inference in the case of inductive-statistical explanation) from empirically well-confirmed statements.[7]

Hempel's covering-law account is a natural consequence of incorporating scientific explanation into a logical empiricist framework. With this account we get a reconceptualization of what it is to achieve scientific understanding. Hempel provides the following formulation:

> a [deductive-nomological] explanation answers the question "*Why* did the explanandum-phenomenon occur?" by showing that the phenomenon resulted from certain particular circumstances, specified in $C_1, C_2, \ldots, C_k$,

[6] He does so on the basis of what Sober calls the "subsistence" and "variability" arguments (Sober, 1978). I discuss these arguments in further detail in Chapter 5.

[7] Naturally, this gets the whole picture into trouble, since empirical laws are generalizations that project to unobserved events.

> in accordance with the laws $L_1, L_2, \ldots, L_r$. By pointing this out, the argument shows that, given the particular circumstances and the laws in question, the occurrence of the phenomenon *was to be expected*; and it is in this sense that the explanation enables us to *understand why* the phenomenon occurred. (Hempel, 1965, p.337)

Understanding is a psychological notion; for it to occur there have to be cognitive beings around to understand things. And Hempel's claim that to understand a phenomenon means that the phenomenon "was to be expected" makes it sound at first like Hempel is equating scientific understanding with a kind of *anticipation* of the *explanandum*, a decidedly psychological notion. But Hempel should not be read in this way for his formulation of understanding is derivative of his account of scientific explanation, and as we have seen, psychology plays no theoretical role in what is essential to explanation on this account. Rather, understanding why a phenomenon occurred amounts to knowledge *that* certain "objective" explanatory relations hold, namely, inferential relations between empirically well-confirmed statements and the *explanandum*.[8] To the extent that knowledge requires belief, this account of understanding might entail a minimal psychological commitment, namely, the capacity to represent the contents of belief, although Peirce analyzes belief in terms of what we are disposed to act on (Peirce, 1966).[9] Either way, for Hempel, the real epistemic work where understanding is concerned is in the inferential relations themselves. And if Frege is right, these inferences hold independently of the characteristic features of cognitive agents. Since Hempel makes repeated reference to the importance of "objectivity", call this approach to understanding "objectivism".

[8] See also Craver discussion of "expect" (Craver, 2007, p. 34). For Hempel, expect is more akin to Cartesian certainty; any *rational* cognitive being, regardless of their particular psychological make-up, who is confronted with a deductive-nomological inference cannot doubt that the *explanandum* ought to have occurred given that the premises of this inference are true.

[9] Thank you to Kevin Hoover for bringing this point to my attention.

One important consequence of Hempel's objectivism is that it entails a certain view about the ideals of understanding, namely the having of a *true* and *complete* description of the empirical facts and their logical consequences. Short of these ideals, understanding is incomplete. Hempel describes different ways that explanations can be incomplete. In some cases we will intentionally omit details for the sake of simplicity. In this case, explanations are "partial" or "elliptical". However, in other cases, their will be gaps in our knowledge that need to be filled in. This gives rise to what Hempel calls an "explanation sketch", a notion that will be useful to have in hand later in this Chapter. As Hempel writes:

> A proposed explanation, for example, which is not explicit and specific enough to be reasonably qualified as an elliptically formulated explanation or as a partial one, can often be viewed as an *explanation sketch*, i.e., as presenting the general outlines of what might well be developed, by gradual elaboration and supplementation, into a more closely reasoned explanatory argument, based on hypotheses which are stated more fully and which permit of a critical appraisal by reference to empirical evidence. (Hempel, 1965, p. 424)

The view that emerges is of science as a progressive explanatory enterprise concerned with the "gradual elaboration and supplementation" of explanation sketches by the incremental filling in of gaps in our understanding. However, as is well known, the covering-law approach to explanation faces serious problems. Consider briefly, by way of example, two problems that have received significant attention in the literature on explanation, namely, the problems of explanatory relevance and of the asymmetry of explanation.

First, imagine, as in Kyburg's example (Kyburg, 1965), we observe that a particular substance dissolves in water. We are informed that what explains its dissolving is that the substance is salt, that the salt has been hexed, and that it is a general law that hexed salt dissolves in water. The explanation in this case conforms to the deductive-nomological template for good scientific explanation: the *explanandum*, the substance dissolving, is a deductive consequence of empirically well-confirmed initial conditions (the presence of hexed salt) and a general law (that hexed salt dissolves in water). However, the salt's having been hexed is not relevant to explaining why the substance dissolved, so the explanation in this case is not a good one. More generally, the deductive-nomological account does not appear to have the resources to distinguish between explanatorily relevant and irrelevant information in a way that accords with our well-considered judgements about explanatory relevance.

Second, consider Bromberger's flag-pole example (Bromberger, 1966). The length and orientation of a flag pole's shadow is a deductive consequence of statements that describe the position of the sun and general laws concerning the propagation of light. But it is also true that the position of the sun is a deductive consequence of statements that describe the length and orientation of the flag pole's shadow and the same laws. Both arguments are deductively valid arguments that conform to the deductive nomological template for good scientific explanation, and yet only the former argument is a genuine explanation; flag-pole shadows do not *explain* anything about where the sun is at any particular moment. The example is meant to show that there is an asymmetry to explanation that the deductive nomological account cannot accommodate.

The problems of explanatory relevance and asymmetry arise as a consequence of the idea that explanations are arguments. In the case of explanatory relevance, since adding premises to a valid argument does not affect the argument's validity, it does not seem possible, given only the resources of deductive-nomological theory, to distinguish between premises that are relevant to explaining the *explanandum* and premises that are not.[10] Furthermore, the inferential relations between statements in

[10] Interestingly, Hempel writes:

a deductive argument are symmetrical in a way that explanation seems to preclude. Explanation and so our understanding of a phenomenon does not appear to depend merely on what inferences we can make from our existing store of empirical knowledge. Rather, something more is required—something, for example, that can satisfy the joint criteria of explanatory relevance and asymmetry. Many philosophers have argued that knowledge of the *causal* processes involved in producing a phenomenon to be explained is the missing link. And with good reason: explanatory relevance appears to track causal relevance in a wide range of cases (e.g., the examples described above); and furthermore, the causal relation is an asymmetrical relation, so it can accommodate the asymmetry of explanation. Of course, positivistically-oriented philosophers like Hempel were, following Hume, their intellectual forbear, deeply skeptical of relying on causation in a respectable epistemology.[11] But logical empiricist epistemology, facing serious challenges, fell out of favor, and with it the constraints on the potential sources of objectivity in theories of explanation. The result is a metaphysical turn in the philosophy of explanation characterized by an *ontological* orientation toward the sources of objectivity. On this approach, explanations are not arguments, but rather objective features of the world.

4.3.2 Salmon on scientific understanding

Salmon's causal theory of explanation is characteristic of the metaphysical turn in the philosophy of explanation and so, in important respects, it contrasts sharply

> reliance on general laws is essential to a D-N explanation; it is in virtue of such laws that the particular facts cited in the explanans possess explanatory relevance to the explanandum phenomenon. Thus, in the case of Dewey's soap bubbles, the gradual warming of the cool air trapped under the hot tumblers would constitute a mere accidental antecedent rather than an explanatory factor for the growth of the bubbles, if it were not for the gas law, which connect the two events. (Hempel, 1965, p. 337)

[11] Salmon notes that Hempel and Oppenheim initially equate general laws with causal laws; however, Hempel changes his view in later work conceding that not all general laws are causal laws.

with the Hempelian approach. However, what remains the same in the move from Hempel's covering-law account to Salmon's causal-mechanistic theory is the commitment to an objectivist conception of scientific understanding. In this section, I describe how Salmon extends this approach to scientific understanding to his own causal-mechanistic model, which remains a (if not the) dominant approach to explanation in contemporary philosophy of science.

Scientific understanding, according to Salmon, "has both practical and intellectual value." The practical value of scientific understanding is "obvious", he writes: "We want to explain why bridges collapse to discover how to prevent such occurrences in the future. We want to explain why certain diseases occur in order to find out how to cure them." (Salmon, 1998, p. 80) However, while scientific understanding is instrumental towards various pragmatic ends, the practical value of scientific understanding is just a side-benefit, for it also has, according to Salmon, "intellectual value", which is something we seek for its own sake. What then, according to Salmon, does scientific understanding consist in?

As in the case of explanation, there are different senses of "understanding." Salmon contrasts *scientific* understanding with three other senses: empathetic understanding ("My wife doesn't understand me", being "the eternal complaint of husbands", and "the standard 'line' for those who intend to be wayward"), understanding of meanings (of language, symbols), and understanding of human behavior in psychology and the social sciences (teleological understanding). A distinctive feature of these other senses of understanding is that they all depend for their applicability on certain psychological characteristics insofar as they involve notions like comprehension, intelligibility, or a psychological "sense" of understanding. However, on Salmon's view, as with Hempel, these psychological notions do not figure into an account of *scientific* understanding. On this point, Salmon is adamant:

> One point that deserves strong emphasis is the absolutely fundamental distinction between "understanding" in the scientific sense and "understanding" in the psychological sense. Understanding in the scientific sense involves the development of a world-picture, including knowledge of the basic mechanisms according to which it operates, that is based on objective evidence—one that we have good reason to suppose actually represents, more or less accurately, the way the world is...This kind understanding may be psychologically satisfying or psychologically discomforting; regardless, the intellectual value remains. Psychological understanding in the empathetic sense may be pleasant and comforting, but it lacks the objective basis furnished by scientific investigation of the world. (Salmon, 1998, p. 90)[12]

What makes scientific understanding distinctive is that it has an "objective basis"—it concerns the kind of knowledge "furnished by scientific investigation of the world", which is to say, the kind of knowledge that scientific explanations provide. Thus, according to Salmon, while explanations may have value because they produces understanding, what genuine "scientific" understanding consists in is just the kind of knowledge that good scientific explanations provide—understanding *just is* explanatory knowledge. Therefore, on Salmon's view, the answer to the question wherein lies the value of scientific explanations, what reasons there are for seeking them, just

[12] In a similar vein, Salmon writes elsewhere of psychological approaches to explanation:

> Whether we have successfully explained the phenomenon, on this view, depends entirely upon whether we have overcome our psychological uneasiness and can feel comfortable with the phenomenon that originally set us to wondering.
>
> the view that scientific explanation consists in release from psychological uneasiness is unacceptable for two reasons. First, we must surely require that there be some sort of *objective* relationship between the explanatory facts and the fact-to-be-explained...Second, not only is there the danger that people will feel satisfied with scientifically defective explanations; there is also the risk that they will be unsatisfied with legitimate scientific explanations.
>
> The psychological interpretation of scientific explanation is patently inadequate. (Salmon, 1984, p.13)

devolves to the question what account or accounts of explanation are the rationally defensible ones, where one of the key criteria that makes an account rationally defensible is that it have an “objective basis”, which is to say, that it be independent of the psychological characteristic of individuals. Salmon identifies two “forms” of scientific understanding that satisfy this criterion:

> The first of these involves understanding our place in the world and knowing what kind of world it is. This kind of understanding is cosmological. The second involves understanding the basic mechanisms that operate in our world, that is, knowing how things work. This kind of understanding is mechanical. (Salmon, 1998, p. 81)

Understanding of the first kind involves having “a unified world picture and insight into how various phenomena fit into that overall scheme” (Salmon, 1998, p. 89); it consists in knowledge of how a phenomenon fits (or can be subsumed) into “the causal nexus.” Understanding of the second kind involves “knowledge of how things in the world work, that is, of the mechanisms, often hidden, that produce the phenomena we want to understand.”(*Ibid.*) In both cases, explanatory knowledge is knowledge of certain objective features of the world, namely, the causal relations between things (e.g. entities, events, etc.) that give rise to the phenomena we wish to understand. Of course, to the extent that explanatory knowledge is *knowledge*, it requires a minimal commitment to certain cognitive features, namely, whatever capacities are required to possess, for example, the belief that certain objective features of the world are related in a particular (explanatory) fashion. But beyond this minimal psychological commitment, the real work, the epistemologically relevant work, is not psychological; it is ontological. Salmon thus provides an ontological alternative to Hempel’s logical criterion. Scientific understanding is not inferential knowledge. It is, according to Salmon, knowledge of the network of objective causal relations that there are and of how the phenomena that interest us fit into that causal network.

Salmon makes this transition from a logical to an ontological criterion explicit by introducing the distinction between the epistemic and ontic conceptions of explanation that I discussed in Chapter 2.[13] According to Salmon, Hempel's covering-law account exemplifies the epistemic conception: on the covering-law theory, the explanation relation is an inferential relation (deductive in the case of deductive-nomological explanations, inductive in the case of inductive-statistical explanations), and inference relations are epistemic relations.[14] By contrast, the ontic conception construes the explanation relation not as an inferential relation between sentences, but rather a relation that fits "the explanandum-event into a discernible pattern" (Salmon, 1984, p. 17). Rather than consisting in knowledge *that* a certain inference holds, explanatory knowledge consists in *how* "a phenomenon fits into the causal nexus", or how it works. So, the task of understanding the natural world amounts to discovering the objective explanatory relations that there are and describing them as best we can.

This distinction between the epistemic and ontic conceptions of explanation has occupied a prominent place in the literature on explanation since Salmon made it. However, in the context of the present discussion, the distinction has very little bearing on what I take to be a central feature of the conception of scientific understanding that Hempel and Salmon both endorse. For although we see a shift in the *source* of objectivity, from logic to ontology, on which Hempel and Salmon ground their

[13] As previously noted, Salmon introduces a third "modal" conception, according to which the relation between *explanans* and *explanandum* is one of "physical necessity" (Salmon, 1984). I do not discuss this conception here.

[14] Salmon writes,

> This explanation could be described as *an argument to the effect that the event-to-be-explained was to be expected by virtue of the explanatory facts*. The key to this sort of explanation is *nomic expectability*...Nomic expectability as thus characterized is clearly an epistemological concept." (Salmon, 1984, p.16)

respective accounts of explanation, what remains constant through this transition is the commitment to objectivity itself, where objectivity is understood as a commitment to a notion of explanatory knowledge that does not depend on psychology except in the minimal sense described. Salmon's ontic approach to explanation is a species of this approach to understanding. But so is Hempel's "epistemic" approach. Thus, the accounts of Salmon and Hempel, despite their significant differences, are characteristic of a general sort of approach to the nature of scientific understanding, i.e. "objectivism". On this approach, scientific understanding is knowledge of certain objective explanatory relations, where those relations are rationally defensible as explanatory relations independently of any appeal to the psychology of those who come to know about them. On this view, psychologistic notions such as comprehension, intelligibility, empathy, satisfaction, or any other psychological "sense" of understanding play no theoretical role in the epistemologically important aspects of scientific understanding.[15] For both Hempel and Salmon, the acquisition of explanatory knowledge amounts to employing various methods for discovering, and then describing, what the objective explanatory relations actually are.

4.3.3 Craver on understanding the brain and explanatory completeness

As discussed in Chapter 2, Carl Craver is firmly in the causal-mechanistic camp where explanation and understanding in the cognitive neurosciences are concerned. His approach to mechanistic explanation is an explicit adoption of Salmon's causal-mechanistic explanatory framework.[16] Thus Craver's mechanistic approach amounts to a commitment to objectivism applied to understanding in the cognitive sciences and so is pertinent to the question as to the nature of understanding in this disci-

[15] Khalifa argues that an explication of understanding is exhausted by theorizing about explanation (Khalifa, 2012).

[16] Though he differs in his account of causation, rejecting Salmon's mark transmission account in favor of a Woodwardian interventionist account.

pline. Furthermore, Craver discusses in detail the theoretical consequences of this commitment, consequences that have significant ramifications for assessing whether objectivism is a plausible approach to understanding in the cognitive sciences. In this Section, I describe how Craver's mechanistic approach falls into the objectivist tradition and the consequences of this approach with respect to the nature of understanding in the cognitive neurosciences.

Recall that according to Craver, cognitive neuroscience explains by discovering and describing mechanisms, where mechanisms are "entities and activities organized such that they exhibit *the explanandum phenomenon.*" (Craver, 2007, p. 6) This distinctively metaphysical emphasis on the nature of mechanisms, insofar as it refers to "entities" and the causal, spatial, and mereological relationships between them, has its roots in the causal-mechanistic account of explanation that Salmon defends. Indeed, Craver adopts at the outset Salmon's "ontic" view as "the correct starting point in thinking about the criteria for evaluating explanatory texts in neuroscience." (Craver, 2007, p. 27) And as I have argued, the ontic conception of understanding is a species of objectivism. To elucidate his commitment to the ontic conceptions, Craver distinguishes between *explanatory texts* and *objective explanations.* Explanatory texts are "descriptions, models, or representations" of objective explanations, which are not representations, but rather "full-bodied things", i.e. in Craver's case, they are *mechanisms.* Explanatory texts, as representations, can be right or wrong, accurate or inaccurate, complete or incomplete. But objective explanations, as things-in-the-world, are not like this. As Craver writes: "There is no question of objective explanations being "right" or "wrong," or "good" or "bad." They just are." (Craver, 2007, p. 27)

Craver's description of the ontic approach to explanation maps nicely onto the objectivist framework of understanding. Adopting Craver's preferred terminology in the context of objectivism, scientific understanding amounts to the having of a representation of an *explanatory text*; genuine scientific understanding is then the having of a representation of a *good* explanatory text. This conception of understanding involves only a minimal commitment to the psychological characteristics of individuals, namely, the capacity to represent explanatory texts. The real philosophical work where understanding is concerned is in assessing whether the explanatory text that one represents is a good one. Craver identifies two normative criteria for good explanatory texts.

First, good explanatory texts are *accurate*, that is, "they correctly represent objective explanations." (Craver, 2007, p. 27) Specifically, good explanatory texts are "how-actually" rather than "how-possibly" descriptions, where mechanistic descriptions are "how-actually" descriptions if they "describe real components, activities, and organizational features of the mechanism that in fact produces the phenomenon." (Craver, 2007, p. 113)[17]

Second, good explanatory texts are *complete*, which is to say, they "represent all and only the relevant portions of the causal structure of the world." (Craver, 2007, p. 111) Craver employs the Hempelian notion, discussed above, of an "explanation sketch" as a contrast to complete mechanistic explanations. A mechanism sketch "characterizes some parts, activities, or features of the mechanisms's organization, but it leaves gaps." (Craver, 2007, 113) These gaps can be problematic because they can impede explanatory progress by creating the "illusion of understanding." For example, Craver describes what he calls "filler terms", which "give the illusion that the explanation is complete when it is not."(*Ibid.*) Common filler terms employed in the neurosciences, according to Craver, include terms like "encode", "inhibit", "modulate", "store", and "represent". These terms "indicate a kind of activity in a mechanism without providing any detail about exactly what activity fills that role."

[17] Hempel discusses the formulation of how-possibly explanations in the context of pragmatic aspects of explanation (Hempel, 1965, pp. 428-30). However, here they are contrasted with "why-necessarily" explanations. Hempel attributes the distinction to Dray.

As with Hempel, making explanatory progress involves the "gradual elaboration and supplementation" of an explanation sketch, filling in gaps by accurately describing the mechanisms that perform the activities (encoding, inhibiting, modulating, etc.) associated with the various filler terms. Between the extremes of a complete mechanistic explanation and a mere mechanism sketch "lies a continuum of *mechanism schemata* whose working is only partially understood."[18] According to Craver, explanatory progress in neuroscience thus "involves movement along both the possibly-plausibly-actually axis and along the sketch-schema-mechanism axis." (Craver, 2007, p. 14)

With this framework in hand, we can formulate the notion of an ideal of understanding that corresponds to what Railton termed an "ideal text" (Railton, 1981). In the case of mechanistic explanations, ideal understanding amounts to representing an ideal text that is an accurate and complete representation of the mechanism (the objective explanation) responsible for the *explanandum* occurring. Of course, achieving understanding in the form of an ideal explanatory text is impossible for many phenomena simply in virtue of their complexity. For example, it would be impossible to provide a *complete* mechanistic model of the brain, a model that includes "all and only the component entities, activities, properties, and organizational features that are relevant to the multifaceted phenomenon to be explained." However, complete understanding serves as, in Kantian terms, a "regulative ideal" against which the practice of cognitive neuroscientists can be measured. To the extent that neuroscientists wish to make explanatory progress, they ought to engage in investigative strategies that approximate this ideal.[19]

[18] We might find that *no* mechanisms performs the activity specified by a filler term, in which case we would have to consider that our explanation sketch fails the accuracy criterion. Research on memory is a nice example and it speaks to the importance of "characterizing the *explanandum*", which Craver discusses at length.

[19] As Craver writes:

> The regulative ideal is that constitutive explanation must describe all and only the

4.4 The case against objectivism

In Chapters 2 and 3, I constructed a case against mechanism as an adequate account of the nature of explanation in cognitive neuroscience, and by extension as an account of the explanatory promise of that discipline. Arguments for mechanistic imperialism in the cognitive sciences are not persuasive. Furthermore, mechanism does not adequately describe the actual practice of the discipline, a discipline we nevertheless have good reason to suppose has significant explanatory promise. However, the failure of the mechanist program in the cognitive sciences is, I believe, symptomatic of a deeper problem, namely, the inadequacy of the objectivist approach to scientific understanding that Craver inherits from Salmon, which in turn is rooted in Hempel's anti-psychologism. The remainder of this Chapter is concerned with advancing an argument to this effect. First, objectivism's commitment to the twin ideals of accuracy and completeness are contradicted by the practice of constructing idealized and abstract models. Second, more generally, objectivism marginalizes as non-essential certain pragmatic ends that are an important component of the ends we seek when we seek to understand a phenomenon. Finally, it is not at all clear that the ideals of objectivist understanding are ends that we *ought* to value for their own sake, and that, for this reason, they ought to guide our best scientific practice.

4.4.1 Idealization and Abstraction

In addition to the explanatory variety characteristic of cognitive neuroscience, explanatory models in this discipline incorporate idealizations and abstractions, where these are broadly understood as the intentional misrepresentation of certain features of a system, and omission of details, respectively, in the construction of explanatory

component entities, activities, properties, and organizational features that are relevant to the multifaceted phenomenon to be explained." (Craver, 2007, p.111) models (Jones, 2005). This has important ramifications for objectivism. Specifically, it is in virtue of, not in spite of, incorporating idealizations and abstractions that models in cognitive neuroscience succeed as explanatory models. This challenges the descriptive adequacy of an objectivist account of understanding in the cognitive sciences—the tendency toward idealized and abstract models in cognitive neuroscience is in direct conflict with the objectivist ideals of accuracy and completeness. If that is right, then something has to give: either cognitive scientists, insofar as they pursue the explanatory strategies that they do, are not doing what they ought to be doing (either on the face of it, or as methodological ideal) and the appearance of progress is an illusion (perhaps encouraged or sustained by our psychological sense of understanding, which can be notoriously misleading), or the conception of understanding that dominates the philosophy of science is problematic in a fundamental way (in certain respects, it is the *opposite* of what scientists are actually doing).

Idealized models in cognitive neuroscience

Idealized models are models that incorporate "useful fictions" or "simplifying falsehoods" (Kaplan and Bechtel, 2011). The philosophical literature on idealization attributes different functions to this practice. One, which Robert Batterman associates with what he calls the "traditional view", is to make models more computationally tractable due to their complexity (Batterman, 2009). By making simplifying assumptions in the early stages of inquiry, scientists are able to get a handle on a complex problem with the idea that, as research progresses, more accurate detail can be incrementally added to the model. This form of idealization is justified on purely pragmatic grounds (Weisberg, 2007a) and it is consistent with the objectivist ideals of understanding—indeed, idealizations function as means to expedite progress in the direction of that ideal.[20]

An alternative is idealization that involves intentionally simplifying all but the most dominant or salient features of a system in the construction of models.[21] It is in virtue of highlighting these features, several philosophers argue, that scientists construct models that promote understanding of a phenomenon under investigation. As Batterman writes of this practice in physics,

> in some cases (and actually in many cases) idealized "overly simple" model equations can better explain and characterize the dominant features of the physical phenomenon of interest. That is to say, these idealized models better explain than more detailed, less idealized models." (Batterman, 2009, p. 429)

That is, on this view, idealizations are incorporated specifically for the purpose of explanation rather than, as in the traditional view, promoting understanding by putting scientists on the right path towards the objectivist ideal. For this reason, call this type of idealization "explanatory idealization".[22] Whether and how idealization serves an explanatory function, and so promotes scientific understanding, depends both on how we interpret the notion of a dominant feature of a system and on supposing that highlighting such features actually promotes understanding. Batterman argues that, given that it is the patterns in nature that occupy our explanatory interests, the dominant features of the system are those that are explanatorily

[20] Weisberg calls this type of idealization "Galilean idealization" in homage to Galileo's use of both theoretical and experimental idealizations "in order to get traction on the problem" (Weisberg 2007, p.641). McMullin describes and defends Galileo's use of idealizations (McMullen, 1985).

[21] Batterman calls this the "non-traditional view"; Weisberg "minimalist idealization". Levy and Bechtel also define idealization as the incorporation of falsehoods that "nevertheless expedite analysis and understanding." (Levy and Bechtel, 2013)

[22] Weisberg identifies a third type of idealization he call "multiple-models idealization", in which multiple, inconsistent models are constructed, "each of which makes distinct claims about the nature and causal structure giving rise to a phenomenon." (Weisberg, 2007a, pp. 645-6) However, I do not discuss this here.

relevant to the occurrence of those patterns, i.e. they explain the repeatability of a phenomenon of interest. Phenomena are repeatable (they constitute patterns in nature) even though many details change from occurrence to occurrence. So, by idealizing or simplifying those features that are variable between occurrences of a repeatable phenomenon, scientists are able to construct models of a system that are stable representations despite changes in details that would otherwise obscure the pattern (Batterman, 2009, p. 430). In a similar vein, Weisberg, following Strevens (Strevens, 2008), characterizes dominance in terms of (causal) difference-making: by idealizing features that are not difference-makers, modelers are able to emphasize the explanatory role of the actual difference-making features. The general picture we get is that explanatory idealization works by simplifying irrelevant features of a system, understood relative to ones explanatory aims. In this way, scientists construct models that promote understanding in a way that more accurate models do not.

It might seem like the objectivist would have to deny that idealization functions in this way. On the objectivist view, better understanding is always in the direction of more accuracy, not less; of better representation, not (intentional) misrepresentation. However, recall that the objectivist ideal is accurate descriptions of the explanatorily *relevant* features of a system, i.e. those objective relations that bear on the behavior of the system in some important way. And whether scientists simplify features of the system that are not relevant, given ones explanatory aims, would seem to have no bearing on the accuracy of the descriptions of features that are relevant. However, it is not clear that explanatory idealization always functions to simplify non-difference-making features. It might also simplify the central feature of a model. And even though the feature of the model would then no longer refer to any actual feature of the phenomenon under investigation, it might be similar enough that the model is still genuinely explanatory. That said, which view of the function of idealization are the modeling practices of cognitive neuroscientists best described? Are models in cognitive neuroscience idealized in the sense of the traditional view, for the purposes of getting traction on a problem? Or do models in cognitive neuroscience incorporate idealizations for the purpose of promoting understanding?

Defenders of strong mechanism endorse the traditional view. This is to be expected given their commitment to an objectivist view of understanding. This commitment is embodied in the model-mechanism-mapping (3M) constraint on explanation in cognitive neuroscience, discussed in Chapter 2. Recall that, according to this requirement, a model explains a cognitive phenomenon to the extent that it maps variables in the model with "identifiable components, activities, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon", and to the extent that the dependencies among the variables correspond to causal dependencies (Kaplan, 2011). This is because these components and their activities are the objective explanatory relations in virtue of which the *explanandum* phenomenon occurs. Thus, the more accurate the mapping and description of the components and their activities, the better the explanation, and so our understanding of the *explanandum* phenomenon.

Kaplan compares the use of idealizations in models of cognition to Craver's "mechanism schemata", which Kaplan characterizes as "elliptical mechanistic explanations in which known details have been intentionally removed." This echoes Hempel's own discussion of the different ways that explanations can be incomplete. And while Kaplan may conflate idealization with abstraction in making this comparison (idealization is best contrasted with accuracy rather than completeness, although, presumably inaccurate models are also, in a sense, incomplete), his general point still holds, namely, that idealizations do not "jeopardize the explanatory status of a model" because they are just ways of simplifying models to make them more tractable, e.g. for beings of limited cognitive capacities such as ourselves.

This strategy for rationalizing the persistent use of idealized models in the cognitive sciences mirrors Hempel's distinction between the "objective" logical core of explanation and its "merely" pragmatic aspects. On Hempel's view, explanation has an internal logical structure that is independent of the various aims and psychological characteristics of individual scientists. Kaplan makes a similar point: explanatory models are explanatory to the extent that they reflect the objective explanatory relations that there are, namely, those relations that are constitutive of the notion of a mechanism. This ontological core is independent of our representations of those mechanisms incorporating idealizations for purposes of representational simplicity. Idealization, on this view, is consistent with the ideal of understanding being in the direction of accuracy. Given limitations on our cognitive capacities, we will have to make trade-offs in our *descriptions* of mechanisms for pragmatic purposes. But those descriptions are, adopting Craver's own preferred terminology, ultimately descriptions of objective explanations, i.e. the actual mechanisms in the world.

The problem with this rationalization strategy is that it looks like cognitive neuroscientists treat models as explanatory *because* they incorporate idealizations, not in spite of doing so, that the function of idealization is distinctively explanatory rather than "merely" pragmatic. In terms of the above discussion, it looks like cognitive neuroscience constructs models that incorporate explanatory idealizations, and, at least in the case of the neuroscience of decision-making, they have done so with core features of the model. In particular, at the heart of the neuroeconomic model of decision-making is the idea that the brain tracks the relative expected value, as this variable is understood in economic theory, of a chooser's feasible alternatives. But this is an idealizing assumption, one that misrepresents the actual functioning of the brain in certain respects. However, the cost of misrepresenting the system in this way is no match for the apparent explanatory benefit that cognitive neuroscientists have achieved in interpreting the data in this way.[23]

Abstract models

Abstract descriptions contrast with descriptions of concrete particulars: entities, events, states of affairs, and the occurrent relations between them. One way to think about what this means in the construction of models of cognition is in terms of omission of detail (Jones, 2005; Levy and Bechtel, 2013; Kaplan, 2011).[24] On this view, an abstract model of a phenomenon includes only certain details of a system while omitting others. This does not mean that abstract models are inaccurate. While idealized models are "mismatched to reality", abstract models "are poor in detail yet potentially true and accurate." (Levy and Bechtel, 2013, p. 243) Abstraction in this sense is thus a matter of degree: different models will differ in the degree or amount of detail that is omitted.

But omission of degrees of detail is not the only way to think about abstraction. Consider normalization of the value signal in the choice network. The model that describes this process is an information-theoretic model that makes no reference to the populations of neurons that actually instantiate the normalization mechanism. For purposes of understanding certain behaviors, it is sufficient that we have a model of

[23] Consider two additional reasons for thinking cognitive neuroscientists treat idealizations of this sort as explanatory rather than as merely a way of making the system more representationally tractable. First, there is the historical point that the idealization in this case came after the discovery of the mechanisms responsible for choice behavior. By interpreting activity in LIP as tracking expected value, cognitive neuroscience gained significant understanding of human choice behavior that a more accurate description would obscure. Second, scientists respond to recalcitrant data not as falsifying the model, but rather by preserving the idealizing assumptions at the core of the model and attempting to rationalize the data from the perspective of the model. This suggest that there is a strong incentive for preserving rather than eliminating idealizing assumptions in this case.

[24] As Levy and Bechtel write:

> Tersely put, abstraction is the omission of detail. An abstract description includes only some of what could, in principle, be said about its subject matter. It leaves matters open, in certain respects. (Levy and Bechtel, 2013, p. 242)

the information-processing strategy that the brain adopts as an optimizing trade-off between discriminatory power and representational capacity in a limited information channel, a model that applies to any number of systems that have certain characteristic features. In this way, abstract models omit not merely certain *amounts* of detail, but also different *kinds* of detail, for example, mechanistic or structural details, as in this example. That is, abstract models differ not only in *degree* of abstraction, but also in *respects*.[25]

Philosophical analysis of the incorporation of abstractions parallels the discussion of idealization in the previous section. Abstraction understood as omission of degrees of detail contrasts with completeness, an objectivist ideal. If that is what abstraction is, then objectivism has to rationalize this practice on pragmatic grounds: scientists incorporate abstractions for the purposes of computational or representational tractability. On the other hand, several philosophers have argued that abstraction serves a specifically explanatory function. Levy and Bechtel claim that by omitting certain details, for example, of "the properties and specific causal powers of the components" of a system, models increase in generality, and consequently function to show how "the same causal features...play a similar role in diverse systems." (Levy and Bechtel, 2013, p. 259) This generality, in turn, increases the explanatory power of the model.[26] On this view, the function of abstraction is to increase understanding. Again, the question becomes how the practice of cognitive neuroscience compares with these alternatives.

Consider how the brain encodes expected value in the choice network. We want to explain why decision-makers choose randomly once the number of options in a choice set exceeds about eight options. According to the neuroeconomic model of

[25] Levy and Bechtel mention this, but they focus almost entirely on abstraction as omission of degrees of detail. However, respect of abstraction seems the far more interesting kind where explanation is concerned, since respects of abstraction appear to correspond to kinds of explanations.

[26] See also Cartwright (1989).

decision-making, what explains this phenomenon is that in the process of encoding the expected value of the options among the chooser's feasible alternatives, the choice network normalizes the expected value of a given option relative to other options. This normalization function optimizes discriminatory power in a noisy information channel, but at the cost of limiting the number of options that can effectively be encoded in the limited channel. The number of options that can be encoded is thus a function of the following parameters: the number of options in the choice set, the stochasticity of the information channel in which the values are encoded, and the information capacity of the channel. These are all abstract parameters that are characterized independently of the mechanistic details that instantiate them.

If we understand abstraction in terms of omission of details, in this case mechanistic details, then the objectivist has to claim that our understanding of the limitations on choice set size is incomplete without filling in the mechanistic details. This is precisely the strategy that defenders of strong mechanism adopt (Piccinini and Craver, 2011; Kaplan and Craver, 2011). But the only sense in which this seems like an appropriate description of what's going on in this case is in the sense that knowledge of the mechanistic details improves our confidence that the model is an appropriate description of the system. That is, our understanding of the limitations on choice set size is best characterized in terms of the information-processing strategies that the brain adopts to solve a particular kind of problem within the constraints of the system, and not by the mechanistic details of how those strategies get implemented. Instead those details, as I argued in Chapter 2, serve as a robust source of evidence that the model really is explanatory. Furthermore, any attempt to cash in the information-processing model for a purely mechanistic model would seem to only diminish the real explanatory power of the information-processing explanation. So where abstraction is understood as omission of detail, it looks like the objectivist strategy for rationalizing the deployment of abstractions once again runs up against a practice that is at odds with the objectivist ideal.

However, the objectivist might have a way out. I have noted that models in cognitive neuroscience are abstract not only in degree but also in respects. One way to think about this is in terms of the kinds of details that a model *admits* rather than the kinds it *omits*. For example, while we can interpret this information-processing model as omitting certain kinds of detail, namely mechanistic details, it only really counts as omission from an objectivist perspective if the activities constitutive of mechanisms exhaust the objective explanatory relations that there are. If the objectivist admits other types of objective relations as potentially explanatory, such as the abstract information-theoretic parameters I have described, then it seems possible to say that abstract models, where abstractions are understood in terms of respects rather than degrees, are consistent with the objectivist ideal. From this perspective, explaining the decision constraints on effective choice set size requires describing in detail the objective relations (in this case mathematical relationships) between the number of choices and the parameters characteristic of a noisy information channel. And this is just what cognitive neuroscientists have done. We can generalize the point in the following way: on this view, different respects of abstraction correspond to different kinds of explanations. The objectivist can claim that our understanding of a system will progress in the direction of the objectivist ideal to the extent that we can describe as completely as possible those objective (though abstract) explanatory relations. In this way, the objectivist can align her theory of understanding with the practice of cognitive neuroscience so long as she is willing to concede that abstract models do not inherit their explanatory force from the underlying concrete mechanisms that instantiate the abstractions that the model incorporates, that abstract relations are among the genuine explanatory relations that there are, and that knowledge of them, i.e. scientific understanding, is valuable for its "intellectual value alone."

4.4.2 Objective explanatory relations

The persistent deployment of idealizations and abstractions, as these are understood in the philosophical literature, in the cognitive neurosciences appears to conflict with the objectivist ideals of accuracy and completeness as objectivist theories are generally understood (i.e. as restricting the class of objective explanatory relations to, for example, causal relations). However, I have noted that there is a sense of abstraction, abstraction in respect rather than in degree, that appears consistent with the objectivist ideal of completeness after all. This strategy is open to the objectivist as long as she admits different kinds of objective relations as explanatory, i.e. different kinds of "objective explanations".

Objectivist theories of understanding typically restrict the class of objective explanatory relations to a particular subclass of the objective relations that there are. To accommodate the practice in the cognitive sciences of incorporating abstractions, the objectivist has to loosen this restriction to include other objective explanatory relations. But which ones? For example, consider Salmon's own general characterization of the ontic view. On this view, explanatory knowledge is knowledge of how a phenomenon fits into a pattern in nature. But there are a vast array of objective, "real patterns" in nature, patterns that are there independently of whether we can discern them or not (Dennett, 1991). Does the mere fact that a phenomenon can be fit into a pattern in nature, even patterns that are not discernible by us, make that pattern explanatory? If not, then on what basis can objectivism restrict the class of explanatory relations to a subset of the objective relations that there are in a way that is not simply *ad hoc*? The objectivist needs a principled reason for restricting the class of patterns to those patterns, knowledge of which amounts to the intrinsically valuable knowledge that we seek when we seek understanding, knowledge prized for its "intellectual value alone".

Only a subset of the patterns that there are are patterns discernible by us, that are *salient* to us. But the objectivist cannot countenance salience as a criterion for restricting the class of explanatory patterns. For whether a pattern is discernible or salient requires adopting a particular cognitive perspective, or "stance", one that depends on our psychological capacities. An alternative possibility is that the genuinely explanatory patterns are the most basic or fundamental patterns that there are—that is, perhaps metaphysics is the appropriate criterion for restricting the class of explanatory patterns.[27] These would include relations like causation, spatial relations, perhaps mereological relations, relations of logical or mathematical necessity, and nomic relations. But if the objectivist adopts this strategy, then she is back in the position of having to rationalize disagreement between the consequences of her theory and the explanatory practices of the discipline. Consider the kinds of explanation discussed in Chapter 3. This fundamental-relations strategy perhaps makes sense of statistical relations and the mathematical relations characteristic of information-processing explanations. However, it is not clear that this strategy works in the case of other features of information-processing explanations, as well as optimality explanations, and functional explanations. These do not seem to be

[27] Ruben draws on Aristotle's theory of explanation to defend a view of this sort. As he writes,

> Aristotle's technical approach introduces a very special and distinctive idea of explanation, and Aristotle's metaphysics provides the justification for so doing. It may be that this concept of explanation was, to a greater or lesser degree, reflected in ordinary or specialist Greek speech, but whether or not it was, is irrelevant. Its defense is metaphysical, not linguistic.
>
> ...
>
> It may be that the concept of explanation that we actually use is outmoded; it has evolved over a long period of time, and it may reflect erroneous, or even incompatible, beliefs about reality. It may no longer fit what we currently think the world is like. It may be so outdated that conceptual tidying-up is no longer sufficient. If so, concept replacement is the order of the day. If possible, a concept of explanation should be adopted that fits what we think the word is like. How we conceive of what the world is like, what its constituents are and how it works, will justify (at least in large measure) choice of concept of explanation.(Ruben, 1990, p. 85-6)

fundamental metaphysical relations and yet they occupy a central explanatory role in the cognitive neurosciences.

The objectivist ideals of understanding lead to difficult questions when measured against the practice of some of our best and most promising scientific practices. Specifically, it is difficult to reconcile the objectivist ideal of complete and accurate knowledge of certain objective explanatory relations with the modeling practices in the cognitive neurosciences. Scientists presumably engage in these practices because they are instrumental toward achieving certain ends that scientists value where understanding is concerned. But if these ends are not objectivist ends, what are they? To shed light on this question, I turn to a more general discussion of the ends we value when we seek understanding.

4.4.3 The pragmatic value of understanding

The Monty Hall problem is on first consideration a striking example of the plausibility of objectivism. In this case, the explanation from probability theory trumps the tendencies of our sense of understanding, which points many of us in exactly the wrong direction. But the proliferation of explanatory strategies in cognitive neuroscience, together with the apparent difficulty of reconciling these strategies with objectivist ideals, raises the question whether objectivism is an appropriate theoretical attitude, despite its intuitive allure. In light of this, it is worth considering the Monty Hall problem anew.

Consider the different epistemic positions that a person might find themselves in with respect to the Monty Hall problem. On one hand, a person might have a working background knowledge of probability theory and the relevant mathematics for calculating the probabilities that certain events will occur under various kinds of conditions. For a person in this position, the explanation identifies the problem as an instance of a more general kind of problem that probability theory provides tools for solving—once the problem is recognized as an instance of a conditional probability problem, it is a relatively straightforward matter of calculating the relevant probabilities. Call a person in this position an “expert”. On the other hand, a person might have no background knowledge of probability theory. Instead, for this person, the explanation is just a formula that generates an answer to the problem. Call the person in this position a “novice”. What is the difference between the expert and the novice in this case? Do both, in virtue of being in possession of an objective explanation, have genuine understanding of the solution to the problem? It doesn't seem so. Indeed, the principal difference between the expert and the novice in this case appears to be precisely the extent to which they understand the problem and its solution. But this is not really a problem for the objectivist. Of course they are different: the novice's understanding in this case is *incomplete* precisely because she does not possess the same background knowledge as the expert. If the novice knew the axioms of the probability calculus, that Bayes's formula is a theorem of these axioms, and that the Monty Hall problem can be solved using the formula, then (and only then) she, like the expert, would have complete understanding.

However, it is not at all clear that even complete knowledge of the theory is sufficient for understanding. Rather, the important thing about gaining expertise about probabilities is not just knowing the theory but knowing how to *use* the theory to solve problems with certain characteristic features. In other words, expertise involves not only possessing significant background knowledge, it also involves the possession of significant *skill* (de Regt, 2009).[28] Without the skills necessary to solve the problem using knowledge of probability theory, this knowledge is explanatorily inert—that is, the difference between the expert and the novice with a complete background knowledge of probability theory still represents a significance difference

[28] As De Regt writes, “Establishing relations between theories and phenomena...crucially depends on skills and judgment.” (de Regt, 2009, p. 28)

in level of understanding. How significant? After all, the objectivist might insist that skills are "merely" pragmatic aspects of explanation, that the explanatory relationship between the theory (Bayes's theorem) and the problem (the Monty Hall problem) in this case is a direct one and does not depend on a person's level of skill, and so the objectivist "essential core" of understanding is preserved. However, as de Regt argues, skill is crucially involved in drawing the connection between a theory or model, and the *explanandum* phenomenon (*Ibid.*). That is, the connection between model and world depends on there being a person to make that connection, and the making of that connection is what skill amounts to.

The modern tradition in the philosophy of explanation is rooted in a conception of explanation according to which explanations are arguments. And perhaps this has obscured the *epistemic* role of skill in accounts of scientific understanding.[29] However, the role of skill becomes more salient when we think of explanation in terms of constructing models to solve problems about nature and by reflecting on the skill required to use models in the sciences, in both theoretical and experimental contexts. The importance of this kind of skill is evident in Kuhn (if we interpret his notion of a paradigm as a kind of model), who makes skill central to his theory of science. On his view, the progress of "normal science" involves to a significant extent figuring out how to extend the paradigm (understood here as an exemplar) to novel circumstances. But this requires significant skill. Indeed, according to Kuhn, the process of training scientists—increasing their level of understanding to the point that they can contribute to the scientific enterprise—involves training students *how* to apply the exemplar to a range of different problem-types. That is, scientific training is, to a significant extent, an exercise in skill development, not merely knowledge

[29] On the covering law account, to explain a phenomenon requires constructing a deductive or inductive argument from premises concerning the laws of nature and particular matters of fact to the conclusion that the *explanandum* occurred. But philosophers are experts at, if nothing else, constructing arguments, so it is no great surprise that a philosopher as skilled as Hempel would fail to take into account the importance of the skill required to construct such arguments.

acquisition. But if possession of an appropriate skill is a significant component of genuine understanding, then understanding is not objectivist. This is because the possession of a skill is not merely knowledge of a rule (Brown, 1988), but rather involves bringing ones cognitive capacities to bear on solving a problem.[30]

If that is right, it leads us to consider that the cognitive states appropriate to genuine understanding are not merely representational states, that these are not the cognitive states we *value* when we seek understanding. Instead, what is required are, at least in part, cognitive states that embody strategies conducive to achieving certain ends, for example, the capacity to recognize that a problem with certain characteristic features is an instance of a more general class of problem, knowing that this type of problem can be solved using a particular model, and then possessing the skill required to apply the model to that problem.[31]

4.4.4 Complexity and the limits of understanding

There are two justificatory routes to objectivism about understanding. The first is negative and involves rejecting the view that scientific understanding depends in any significant way on the psychological characteristics of individuals. The second is positive and is concerned with endorsing the claim that knowledge of objective relations is knowledge that is valuable for its own sake, so scientists should strive for such knowledge when they seek to understand nature.

The Monty Hall problem is a case where complete knowledge of probability theory is not sufficient for genuine understanding. But it is at least a case where complete

[30] The claim that skills are not simply internalized rules is not uncontroversial. See Reber (1993), Trout (2002).

[31] Another possibility with respect to the pragmatic ends of understanding concerns the division of intellectual labor in the sciences through disciplinization. Models that promote understanding might be those that exploit this disciplinization by identifying distinct problem areas that can be investigated with relative independence even while serving as a guide for what research questions ought to garner attention. In this way, a model would promote a progressive research program across a broad range of disciplines. The neuroeconomic model may be one such example.

knowledge may genuinely be possible. In many domains, cognitive neuroscience among them, complete knowledge of certain objective relations is impossible due to the sheer complexity of the system under investigation. The brain is an enormously complex system, perhaps the most complex system we know of. It would be impossible for any one human being to represent, for example, the complete causal structure of the brain. Worse, no two brains are exactly alike, so there is no single complete causal structure that the human brain exemplifies.

However, scientists have ways of dealing with this complexity. As discussed above, one is to incorporate idealizations and abstractions into their models of the functioning brain. As we have seen, for the objectivist this serves a pragmatic function that is distinct from the core of what understanding consists in—they are techniques for making descriptions of phenomena more tractable given our cognitive limitations. A consequence of this view is that there will be some explanations that human beings will never understand in virtue of the fact that we cannot represent them completely. I have argued that the objectivist account does not adequately describe what we actually value when we seek understanding. Is is nevertheless something that we ought to value even when it can never be achieved, an ideal that ought to guide our best scientific practice? I do not think it is obviously the case that it is.

The question whether objectivist understanding is valuable for its own sake despite being, in many cases, unattainable, echoes, in some respects, the question whether Cartesian certainty is an appropriate epistemic ideal. If certain knowledge is something we can never have, then why is it a kind of knowledge on which to focus our epistemic attentions? If what actually concerns us is the epistemic access that we could possibly enjoy, then it seems like our focus should be on developing an epistemology that reflects what this amounts to and the circumstances under which it is possible. Even so, the intuitive pull of certain knowledge remains if for no other reason than the value we place on the attainment of truths. Even supposing we can never be certain that any of our beliefs are true, we should at least try to get as close as we possibly can.[32] But this points to an important disanalogy between certain knowledge and objectivist understanding. In the case of understanding, if truth, or something close enough, is part of what we are after, we don't need objectivist understanding to achieve it. What is distinctive of objectivism is not its realism, but rather its antipsychologism, and understanding might be both essentially psychologistic and dependent on truth. For example, a model can tell us something about the way the world really is without being either complete or completely accurate. Indeed, I have argued that one of the criteria for claiming that the neuroeconomic model really is explanatory is that we have good evidence that the model tells us something about what's really going on in the case of human decision-making. The question of the relationship between models and the world is a deeply interesting one that I cannot hope to address here. But the point is that whatever that relationship is, it is entirely consistent with an attitude that prizes truth. That is, it is possible to both value truth *and* to think that the epistemic demands of understanding are, at least in part, and to a significant extent, psychological demands.

4.5 Conclusion

Salmon asks why explanations have value, why they are something worth seeking. On his view, explanations have value because of the kind of knowledge they give us, namely, understanding. But understanding, according to Salmon amounts to explanatory knowledge, so the question devolves to what that knowledge consists in. I started this Chapter by discussing one prominent approach to answering this question. According to this approach, adopted by Hempel, Salmon, and Craver, explanatory knowledge consists in knowledge of certain objective relations between *explanans* and *explanandum*, that is, relations that are independent of the psy-

[32] See Stich (1990) for a dissenting view on the value of truth.

chological characteristic of individuals. In the case of the covering-law account, the relation is an inferential relation between statements, which is epistemic. On Salmon's ontologically-oriented view, the relation is one of fit into the causal nexus. In both cases, the explanation relation is construed independently of the psychological characteristics of the individuals who employ it, except in the minimal sense that explanatory knowledge requires the capacity to appropriately represent the explanatory relations that there are.

I do not find compelling the argument that objectivist understanding is what scientists actually seek as a matter of course in attempting to understand cognitive phenomena, nor that it is a kind of knowledge that is valuable for its own sake. Thus, objectivism is on one hand, inconsistent with certain cognitive states we actually value (useful states), and on the other hand, irrelevant to satisfying others that many think we ought to value (truth). In fact, reflection on the practice of cognitive neuroscience suggests that the nature of understanding and the strategies that scientists pursue to achieve it are much richer and more complex than objectivism suggests. In particular, by excising any appeal to our psycho-biological constitution, objectivism misses the essential link between understanding and the kinds of cognitive creatures that we are. In the next chapter, I outline and defend a normative psychologistic account of understanding that better accounts for the ends we actually value and that we ought to value where understanding is our aim.

5

Naturalizing Understanding

5.1 Introduction

In Chapter 4, I criticized a prominent view of scientific understanding, “objectivism”, according to which understanding is just a traditional form of propositional knowledge of certain objective explanatory relations. I argued that consideration of the ends we actually seek where understanding is concerned suggest these ends are, in part, but to a significant extent, pragmatic ends. For example, genuine understanding requires the skill or capacity to apply a theory or model to real-world phenomena. It may also require that explanatory models organize inquiry into progressive research programs. While we value truth as an end of understanding, valuing truth is consistent with rejecting objectivist ideals. Indeed, by emphasizing accuracy and completeness, objectivism may actually obscure the truth about a phenomenon that really matters with respect to our explanatory aims. Thus, I find no convincing rationale for thinking that the objectivist ideals are what we ought to value either.

In this chapter I outline an alternative psychologistic account of understanding, according to which understanding depends in some significant respect on the nature of our shared psychological capacities. More specifically, I defend an account of scientific understanding, according to which achieving understanding about a phenomenon amounts to achieving certain cognitive states that constitute effective strategies for satisfying the ends of understanding, where these ends are, at least in part, but to a significant extent, pragmatic ends.[1] This account of understanding provides a framework for an empirical investigation, broadly construed, into the nature of understanding.

This chapter is organized as follows. First, I build on work by Elliot Sober to characterize different varieties of psychologisms where scientific understanding is concerned. In doing so, I outline the account of understanding I aim to defend. I then defend this account against objections that trace back to Hempel's objectivist approach to explanation and understanding.

5.2 Three psychologisms

Broadly speaking, psychologism is just the denial of objectivism—it is the view that understanding a phenomenon *depends on* characteristics or features of human psychology. But this is a very broad characterization. Specifically, what does it mean to say that understanding "depends on" psychology? Elliott Sober describes

[1] Note that the fact that the ends of understanding are, at least in part, pragmatic ends does not by itself entail psychologism. A possible alternative is an ends-relative pragmatism about understanding, according to which understanding is a matter of pursuing strategies, non-psychologically construed, that effectively lead to the ends of understanding. In Chapter 2 I raised the issue of how to analyze rationality in the context of decision-making. Taking rationality to be instrumental rationality, I concluded that it concerns identifying effective decision-procedures for achieving the outcomes we value and the circumstances under which they are effective. But a major question concerned whether we should think of strategies as effective *per se* or relative to our psycho-biological constraints. To the extent that we are concerned with how *we* can best make decisions, I argued that that the latter is the appropriate theoretical attitude. The general case of understanding parallels these considerations: to the extent that our aim is to assess how we can have understanding, we ought to consider effective strategies toward achieving the ends of understanding in light of our psycho-biological constraints. If that is right, then an ends-relative pragmatism coincides with psychologism for whether a strategy is effective will depend on a characterization of those constraints.

psychologism in epistemology generally as a “family of views, all tending to downplay or deny distinctions between epistemology and logic on the one hand and psychology on the other” (Sober, 1978, pp. 165-6). He distinguishes between three general varieties of psychologisms: metaphysical psychologism, the view that “the laws of logic and the characterization of rationality that epistemology seeks to formulate are *about* human mental activity”; epistemological psychologism, the view that “denies that there is any question of justifying a logical rule or epistemological maxim above and beyond the question of whether it is in fact followed in practice.”; and the view that “the rules of correct reasoning that logicians and epistemologists try to describe have *psychological reality.”* This taxonomy of psychologisms is a useful starting point in the present discussion of understanding.

5.2.1 Metaphysical Psychologism

One way that understanding might depend on psychology is in line with what Sober calls “metaphysical psychologism” about logic and epistemology. According to this view, the “truth-makers” of statements concerning the norms of rational inference are, at least in part, facts about human psychology and not merely facts about the world. Since Frege, the philosophical tradition has generally rejected this kind of idealism where epistemology and logic are concerned. That certain rules of inference and epistemological principles are rational is true (if it is true), not in virtue of facts about human cognition, or the “laws of thought”, but rather because they are, for example, *necessarily* true or because, adopting a naturalist perspective, they constitute effective means for achieving certain desirable goals (for example, making good decisions). In either case, it looks like the truth of the principles of rationality does not depend on human psychology instantiating such principles.

Sober invokes Frege’s “subsistence argument” in defense of this anti-psychologism. As with the truth of mathematical claims, the truth of logical and epistemological maxims is independent of there being any human beings and their characteristic psychologies around to appreciate them. But if metaphysical psychologism is true, then the truth of logical and epistemological maxims does depend on the features of human psychology. So metaphysical psychologism is not true of logic and epistemology. Rather, the principles of logic and epistemology are objective truths. Whether human psychology actually instantiates those principles, whether they have, as Sober puts it, “psychological reality”, is a different question altogether. For, supposing they do have psychological reality, it would not be facts about our cognitive economy that makes it true that they are sound principles. Rather it would simply be the case that sound principles are, to our great epistemic fortune, instantiated in the cognitive machinery that constitutes human psychology.

But suppose, as naturalism does, that rational principles are those principles that reliably lead to instrumentally desirable outcomes. If they are effective at achieving such outcomes for us, then they are effective within the constraints of our particular psycho-biological constitution. If that is right, then whether it is true that a principle is rational will depend on a characterization of the (perhaps distinctive) constraints embodied in human cognition—that is, it will depend on human psychology. While this is not the full-blown metaphysical psychologism according to which, e.g. “the laws of logic just are the laws of thought”, it is instructive where understanding is concerned.

Metaphysical psychologism about understanding is the view that what makes it true that one understands a phenomenon is, at least in part, features of human psychology. If understanding is a psychological state or cognitive capacity that is conducive to achieving certain desirable ends, then the truth of whether one understands a phenomenon will depend on whether the strategy or means one adopts is effective at achieving the ends desired. As in the case of logic and epistemology generally, one might think that cognitive strategies for achieving the ends of understanding (whatever those happen to be), even supposing they are instantiated by human cognitive machinery, are not good strategies because they are so instantiated, but rather for reasons independent of the psychology we happen to possess. Indeed, we have whole disciplines dedicated to discovering and developing optimal strategies for achieving a wide range of ends in engineering, mathematics, computer science, etc. But the notion that for any given problem there is an ideal solution, one that is independent of any design constraints is problematic. Engineers, when they are solving a problem, do not first develop the "ideal" solution to the problem and then implement that ideal as best they can given constraints on time, resources, the laws of physics, etc. Rather, the problem together with the design constraints are what define the solution space to begin with. In other words, whether a strategy is effective is relative to the constraints within which a problem must be solved. Indeed, the very idea of an "effective strategy" appears to presuppose that certain constraints define the possible solution space.[2]

By way of example, recall the Monty Hall problem discussed in Chapter 4. In particular, consider the difference between the novice about probability theory and the expert. The novice has an answer to the problem of how to maximize her chances of winning the car. But unlike the expert, the novice is in no position to exploit this information to achieve any practical ends. In other words, supposing that the ends of understanding are, at least in part, practical ends, the answer in this case is not an effective strategy for the novice at achieving these ends. How could this be? It must be because something about her psycho-biological constitution is not configured appropriately to achieve these ends. In other words, what makes it false that the novice understands is not that she lacks the objective facts about this case—she has

[2] Example: "Ideal" machine learning strategies are only optimal relative to the constraints of the machines that are doing the learning. Another example: computational strategies are alluring as possible cognitive strategies precisely because we know that such strategies can be instantiated by physical systems, which we take to be a *prima facie* constraint on any theory of mind.

those facts, but rather something about her psycho-biological constitution—she is not in a position to exploit the information she has to achieve the ends of understanding.

If we think of understanding as a design solution to the problem of achieving certain desirable ends (the ends of understanding) then what makes it true that one understands a phenomenon is whether the cognitive strategy one deploys is an effective means (for cognitive beings such as ourselves) of achieving those ends. And this will include both facts about the world and about our psycho-biological constitution. Thus, metaphysical psychologism should not be dismissed out of hand where explanation and understanding are concerned.[3]

This preliminary defense of metaphysical psychologism has relied on a series of hypotheticals, in particular, on whether naturalism is a defensible approach to inquiry about understanding and whether the ends of understanding are, at least in part, pragmatic ends. I therefore turn to a discussion of these issues.

5.2.2 Epistemological Psychologism

Another way that understanding might depend on psychology has to do with what Sober calls "epistemological psychologism". Where metaphysical psychologism concerns what makes it *true* that the principles of logic and epistemology are rational (i.e. where understanding is concerned it is a view about the truth-makers of claims about understanding), epistemological psychologism concerns the grounds on which we are *justified* in believing they are rational. Specifically, it is the view that justifying a logical or epistemological maxim is not independent of the project of specifying our actual epistemic practices. Two examples of this kind of view to which Sober refers are Quine's claim that given the failures of (Carnapian) foundationalism, epistemology ought to be incorporated as a component of psychology (Quine, 1969),

[3] An example of something that's not an effective strategy: relying on one's sense of understanding, since this is not reliably conducive to achieving ends. What is effective? Constructing models that withstand empirical scrutiny, i.e. are confirmed by evidence.

and Goodman's appeal to reflective equilibrium as a means of justifying inductive inferences, i.e. that doing so is continuous with an analysis of our actual inductive practices (Goodman, 1983).[4] Epistemological psychologism is thus equivalent to the methodological naturalism I discussed in Chapter 2. And Sober rejects this view for familiar reasons: on grounds that naturalized "epistemology" is an unwarranted abdication of normativity (Kim, 1988; Fodor, 1991). This abdication is unwarranted, argues Sober, because justification, and so normativity, is central to anything recognizable as epistemology, and because out-of-favor foundationalisms do not "exhaust the possibilities of epistemology." But, as I discussed in Chapter 2, this mistakes a rejection of traditional justification for justification altogether, and naturalism needn't endorse traditional justification to give normativity its due: while foundationalist conceptions of justification do not exhaust the possibilities of epistemology, it is equally true that traditional notions of justification do not exhaust the possibilities of normativity. Thus, merely from the fact that naturalism rejects traditional notions of justification, it does not follow that it constitutes an "abdication" of norms. Instead, rational (justified) inferences or epistemological principles are just those processes or capacities that human beings employ that reliably lead to *success* under some reasonably well-defined range of circumstances.

By extension, epistemological psychologism about understanding is the view that we are justified in thinking that understanding is genuine when the psychological states we achieve or the cognitive capacities that we employ are consistent with strategies known to be effective strategies for achieving the ends of understanding, where those ends are ends that are desirable for the kinds of cognitive creatures that we are. In this way, inquiry into the nature of understanding is continuous with inquiry into those cognitive capacities that human beings deploy in attempting to

[4] Goodman writes: "A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend." (Goodman, 1983, p. 64)

understand various phenomena. This does not mean that it is a merely descriptive enterprise. Whether our claims to having achieved understanding are justified will depend on whether the cognitive capacities we have deployed to achieve it are conducive to the ends of understanding. Thus, to the extent that the traditional approach (objectivism) faces problems, and to the extent that methodological naturalism is an adequately normative approach to inquiry about understanding (and so possesses that feature that make it recognizable as epistemology) it is, contra Sober, a defensible position.[5] The important point to take away from this discussion is that consideration of what a commitment to epistemological psychologism entails provides a framework for inquiry into the nature of understanding. Specifically, it involves three "dimensions" of inquiry organized around the following guiding questions:

1. What are the ends of understanding?

2. What is the nature of those cognitive capacities for understanding that we deploy in attempting to achieve those ends?

3. Under what conditions is the deployment of those capacities an effective strategy for achieving those ends?

Epistemological psychologism further entails certain preliminary answers to these questions. Consider the question of the ends of understanding. As I have argued, epistemological psychologism is differentiated from an uncritical descriptivism on grounds that what constitutes justified practices where understanding is concerned is that these practices are conducive to success. Thus, the ends of understanding are understood in terms of whatever success consists in. Exactly what those ends are

[5] Of course, naturalism is subject to a question-begging argument just as objectivism is with regards to the success criterion where understanding is concerned. In this case, the objectivist can reply that naturalism begs the question against objectivism by assuming that success is empirical success, whereas for the objectivist, what constitutes success is precisely the question at issue. But the objectivist is in no better position on this score than the naturalist. And objectivism faces the descriptive adequacy challenge that is the subject of Chapter 4.

is, to a significant extent, an empirical question that will be answered by drawing on a broad range of empirical considerations, to include consideration of what our cognitive capacities for understanding are actually conducive to achieving. The ends might include, as I have argued, pragmatic ends, for example the capacity to apply a model toward solving problems about nature. But the ends might also include truth or something close enough, so long as there is room in our account of empirical success for turning data into evidence that a model tells us something about what is really going on.

5.2.3 The Psychological Reality of Effective Strategies

Given some idea of the ends of understanding we can hypothesize about the best means to achieve those ends given what we know about our own psycho-biological constitution. The question becomes to what extent our actual cognitive systems match our theories about what those strategies might be, whether the cognitive models we construct have, as Sober calls it, ”psychological reality”. As a brief illustration of how this would work, consider three examples of cognitive models that have been proposed as contender psychologies of understanding.

The sentential model The model of cognition that dominated much of twentieth century philosophy has its roots in developments in logic and language. This tradition achieves full fruition in the classical computational model of mind, according to which cognition involves computational transformations of strings of symbols, where these symbols represent meaningful sentences, and the rules of transformation among these sentences are the rules of logical inference. On this view, knowledge is propositional knowledge.

The deductive-nomological model of explanation fits very well with such a view. We represent empirically well-confirmed statements that refer to general laws and particular matters of fact (initial conditions). We can then infer the deductive consequences of these statements. According to Hempel, explaining a phenomenon (in the case of the deductive-nomological account) just amounts to being able to deductively infer a statement of the phenomenon's occurring from these observationally well-confirmed statements. Genuine understanding of a phenomenon is then just knowledge of this inferential process—it amounts to knowing that the explanandum *was to be expected* (in virtue of being a deductive consequence) given empirically well-confirmed statements.[6]

Churchland's PDP model The sentential model is not uncontroversial. For example, Paul Churchland challenges the psychological plausibility of the model and so rejects the deductive-nomological approach to explanation that relies on it (Churchland, 1989). In naturalistic terms, if a model is psychologically implausible, then it could not be an effective strategy for cognitive beings such as ourselves for achieving the ends of understanding, so it could not be what genuine understanding consists in. Instead, Churchland defends a neurophysiological account of explanatory understanding, according to which,

> Explanatory understanding consists in the activation of a particular prototype vector in a well-trained network. It consists in the apprehension of the problematic case as an instance of a general type, a type for which the creature has a detailed and well-informed representation. (Churchland, 1989)

On this view, the variety of explanatory strategies one sees in, for example, the cognitive sciences, can be incorporated into a unified account of explanation, where "what differs is the character of the prototype that is activated." (Churchland,

[6] Sober defends the "psychological reality" of this view (Sober, 1978).

1989, p. 212) Churchland provides several examples of these prototypes, including "property-cluster prototypes" (corresponding to taxonomic explanations), "etiological prototypes" (corresponding to causal explanations), "practical prototypes" (corresponding to functional explanations), and "motivational-prototypes" (corresponding to belief-desire explanations). On this view, explanatory understanding is a species of *recognition*—one classifies a phenomenon of interest as an instance of a general type, and in doing so, one gains a massive amount of information about the explanandum otherwise unavailable. As Churchland writes, prototype activation:

> represents a major and speculative *gain* in information, since the portrait it embodies typically goes far beyond the local and perspectivally limited information that may activate it on any given occasion. That is why the process is useful: it is quite dramatically ampliative. (Churchland, 1989)

One point worth noting is that although Churchland emphasizes neurobiology in developing this account, most of the theoretical work is done at the level of the type of prototypes that are instantiated by, according to Churchland, these massively parallel connectionist networks of neurons. Given that, our capacity for understanding is best characterized from a more cognitive, and less "mechanistic" perspective. If we think of vector-prototypes in more general terms, namely, as cognitive models of a particular system, then we can focus our investigatory attention on the nature of these models.[7] The model-model of understanding does just this.

The model-model of understanding The importance of modeling in science has been a subject of significant recent attention in both the history and philosophy of science.

[7] In Chapter 1, I argued that cognitive neuroscience explains by developing cognitive models and marshaling evidence, in the form of mechanistic discovery, in favor of those models. Churchland can be seen as doing something very similar here: proposing a cognitive model of understanding, according to which understanding amounts to subsuming a phenomenon under a particular representation (as Craver characterizes the view) and providing evidence in favor of such a model in the form of the connectionist theory he favors.

For example, models can be seen as playing a prominent role in Kuhn's characterization of scientific practice as being organized around a "paradigm" (Kuhn, 1996); Nersessian provides several detailed accounts of the importance of models in the history of science (Nersessian, 1984, 1995, 2006); Cartwright gives a prominent epistemic role to the construction of models in physics (Cartwright, 1983); Weisberg explores the nature and function of modeling (Weisberg, 2006, 2007b, 2013); and Waskan defends an account of understanding according to which understanding a phenomenon involves having what he calls an "intrinsic cognitive model" of the mechanisms responsible for a phenomenon (Waskan, 2008).[8] One natural question that comes out of this work is whether our cognitive capacity for understanding might consist in the construction and deployment of cognitive models of phenomena of the kind Waskan hypothesizes, i.e. whether models have "psychological reality".

5.3 Two Objections to Psychologism

As Waskan notes, anti-psychologism about understanding is motivated on both metaphysical and epistemological grounds (Waskan, 2011). The metaphysical argument purports to show that what makes claims about understanding true are only objective matters of fact, and not the psychological characteristics of individuals. The epistemological argument purports to show that claims about understanding are justified independently of facts about our psychology, that is, on whether understanding satisfies certain philosophical norms (e.g. the explanans deductively entails the explanandum, or the explanans is causally relevant to producing the explanandum).

[8] Waskan writes:

> to have an [inanimate] explanation is to have the belief that a certain mechanism is, or may be, responsible for producing some happening, where such beliefs are constituted by mental representations [viz., intrinsic cognitive models] of those mechanisms. It is largely in virtue of our awareness of the information conveyed by these representations that events and physical regularities are rendered intelligible (Waskan, 2008).

While I have attempted to show that psychologism is a plausible thesis, if psychologism is to be defended as a viable alternative to objectivism, it ought to be able to respond directly to these objections. Furthermore, consideration of these objections will allow us to make progress in filling in the investigative framework I have proposed. My principal focus shall be on Hempel's own development of these objections as I take it that his treatment of these objections is representative of what continues to motivate anti-psychologism about understanding.

5.3.1 The metaphysical objection

The metaphysical objection is an extension of Frege's subsistence argument from logic and epistemology to the domain of explanation. If understanding is psychologistic in the metaphysical sense, then the truth of understanding claims, at least in part, depends on human psychology. But genuine claims to understanding are *not* relative to the psychological characteristics of individuals: there are objective facts of the matter about what the world is like independently of human beings and their psychologies, and these include explanatory facts—why the world is the way it is is true even if there are no human beings around to appreciate it. Understanding, on this view, merely consists in the description of these objective explanatory facts; the more detailed and accurate the description, the better our understanding of the phenomenon in question. Therefore, understanding is not psychologistic in the metaphysical sense. For example, on Hempel's deductive-nomological account, what explains a phenomenon is that it is a deductive consequence of statements that describe one or more empirical laws and the initial conditions. This is true independently of whether there are any human beings around to construct the necessary deductive entailments. On a causal account, a phenomenon is explained by citing its causes, and the causal relations that hold between an *explanans* and the phenomenon to be explained hold independently of the existence of explainers. This is objectivism.

In Chapter 4, I challenged this objectivist view of understanding. In the case of at least one scientific discipline that appears to have considerable explanatory promise, objectivism does not adequately capture what scientists actually seek as a matter of course, nor does it seem plausible that mere objectivist knowledge is valuable for its own sake. In the absence of this descriptive adequacy, objectivism is in no better a position than psychologism with respect to answering what understanding consists in. I have also argued above that there is a perfectly reasonable sense, given understanding is a cognitive state that has certain ends, in which the degree to which understanding constitutes an effective means of achieving those ends will depend on the psycho-biological constraints embodied in our particular cognitive economy. These are sufficient reasons I take it for thinking that metaphysical psychologism has at least initial plausibility. So without some further rationale, the subsistence argument merely begs the question against the defender of psychologism: it just denies that understanding is psychologistic in the metaphysical sense.[9] However, Hempel provides a further rationale. So I shall consider his version of the objection.

In defining metaphysical *anti*-psychologism, Hempel is primarily concerned with defending the deductive-nomological account of explanation against pragmatic theories of explanation, particularly the pragmatic view that Michael Scriven defends.[10] On Scriven's view, the pragmatic aspects of explanation—those aspects that are relative to individuals aims, intentions, desires, etc.—are essential to what explanation actually is. But according to the subsistence argument, to the extent that the pragmatic aspects of explanation involve the psychological characteristics of individuals, they are not essential to the nature of explanation. Hempel challenges

[9] Furthermore, it is not at all obvious that the apparent self-evidence (at least for Frege, Hempel, and Sober) of the objective principles of logic and epistemology extends to explanation and understanding as well.

[10] Other pragmatic accounts include van Fraassen (1980), Garfinkel (1981).

Scriven's pragmatism on precisely these grounds. While we should expect to find that different explanations are appropriate for different explanatory aims, and have different psychological effects on different individuals (given various factors), from this it does not follow that explanation is thereby *essentially* pragmatic. Hempel draws directly on the Fregean point concerning mathematical proof to motivate this point. While there are pragmatic aspects to how individual mathematicians engage in proving mathematical claims, we would not thereby conclude that proof is an essentially pragmatic enterprise; proofs of mathematical claims are valid independently of these pragmatic concerns. The same, Hempel argues, goes for explanation. While there are various pragmatic aspects to how individuals engage with the process of explaining, we should not thereby conclude that explanation is essentially pragmatic. But the question remains whether the analogy is an appropriate one and Hempel must provide some further rationale to show that it is. He does so, following Scriven, by characterizing pragmatic accounts in terms of psychology and "the realm of understanding".[11] According to Hempel, on a pragmatic theory of explanation, "to explain something to a person is to make it plain and intelligible to him, to make him understand it." (Hempel, 1965, p. 425) Explanation, on this view

> requires reference to the persons involved in the process of explaining. In a pragmatic context we might say, for example, that a given account A explains fact X to person P1. We will then have to bear in mind that the same account may well not constitute an explanation of X for another person P2, who might not even regard X as requiring an explanation, or who might find the account A unintelligible or unilluminating, or irrelevant to what puzzles him about X. Explanation in this pragmatic sense is thus a relative notion: something can be significantly said to constitute an explanation in this sense only for this or that individual. (Hempel,

[11] Scriven says: a complete explanation "is one that relates the object of inquiry to the realm of understanding in some comprehensible and appropriate way" (Scriven, 1962).

1965, pp. 425-6)

In this passage we get to the crux of the matter: the pragmatic aspects of explanation, those aspects that make an explanation “plain and intelligible” or *understood*, vary from person to person. But if explanatory knowledge (i.e. understanding) is objective, then it does not vary in this way. So the pragmatic aspects of explanation are not essential to what explanatory knowledge consists in. Hempel concludes that psychologism about explanation is an untenable position.

Note, however, that this latter claim, that *psychologism* is untenable, only follows if psychologism is exhausted by the pragmatic aspects of explanation that Hempel considers, namely, aspects that vary from person to person. But as Michael Friedman points out in a well-known reply to this argument, in drawing this anti-psychologistic conclusion, Hempel appears to equivocate on the meaning of “pragmatic” as meaning *psychological* on one hand and *subjective* on the other (Friedman, 1974). The problem is that, if Hempel means to equate pragmatic with psychological rather than subjective, then the subsistence argument fails. At the core of Hempel’s metaphysical objection to psychologism is the claim that the pragmatic aspects of explanation vary to a significant degree across individuals, which is to say that individuals differ in what they find “intelligible”, or “illuminating”, or what they empathize with, or see as familiar—that is, in what they understand.[12] But not all psychological features are subjective in this way. In fact, our best cognitive science tells us (and indeed must take as a methodological assumption) that the vast majority of our psychologi-

[12] Salmon makes a similar point:

> we must surely require that there be some sort of *objective* relationship between the explanatory facts and the fact-to-be-explained. Even if a person were perfectly satisfied content with an ’explanation’ of the occurrence of storms in terms of falling barometric readings, we should still say that the behavior of the barometer fails objectively to explain such facts. We must, instead, appeal to meteorological conditions.
>
> ...
>
> The psychological interpretation of scientific explanation is patently inadequate.

cal characteristics are characteristics that we share.[13] So from facts about variations in what individuals find “illuminating”, one cannot conclude that understanding is anti-psychologistic. Friedman writes,

> I don’t see why there can’t be an objective or rational sense of ’scientific understanding’, a sense on which what is scientifically comprehensible is constant for a relatively large class of people. Therefore, I don’t see how the philosopher of science can afford to ignore such concepts as ’understanding’ and ’intelligibility’ when giving a theory of the explanation relation.” (Friedman, 1974, p. 8)

As Sober notes, it is a curious fact that Hempel does not consider this possibility. However, notice that in describing the pragmatic aspects of explanation, Hempel emphasizes the *finding* of something as intelligible or illuminating, that is, the *feelings* of understanding.[14] In other words, as Waskan notes, Hempel equates the psychological aspects of explanation with what Waskan calls the “commonly associated phenomenology of explanation”, which is to say the “sense” of understanding. And if *that* is what psychologism says understanding consists in, then skepticism towards a privileged role for psychology is not difficult to appreciate. First, it is not easy to articulate what this “sense” of understanding amounts to. It is variously described as a sense of familiarity (Dray, 1979), a feeling of empathy (Hempel, 1965; Salmon, 1998) an “Aha!” experience or a “seductive feeling of cognitive fluency” (Trout, 2007), or as like an orgasm (Gopnik, 1998).[15] How could something so vague as

[13] Sober makes essentially the same point. He also notes that it is a curious fact that the positivists, including Hempel, did not emphasize this assumption given that the preeminent cognitive science of the day (e.g. behaviorism) was almost exclusively focused on objective psychological laws—indeed it was advanced as a plausible theory for precisely this reason.

[14] Of course, he most likely does so because he is responding to a particular tradition, one that includes Scriven, that places emphasis squarely on such features.

[15] Gopnik suggests this sense has a function, as orgasms have functions. We have to be motivated to seek cognitive states, and we have to have some idea of when we have achieved it.

"Aha!" figure in a philosophical account of genuine scientific understanding? Second, as Hempel notes in the passage quoted above, the extent to which individuals experience this sense varies widely. Finally, the history of science is, from one perspective, a repository of disproven theories, all of which at one time produced a strong *sense* of understanding about the things they were supposed to explain. As Newton-Smith notes, "Once upon a time, many thought that the fact that there were seven virtues and seven orifices of the human head gave them an understanding of why there were (allegedly) only seven planets" (Newton-Smith, 2001). Like Hempel and Salmon before him, Newton-Smith insists that we must "distinguish between real and spurious understanding", and that it is only "a philosophical theory of explanation that will give us the hallmark of a good explanation", and with it the mark of real understanding.[16] Indeed, not only is our sense of understanding a notoriously unreliable indicator of genuine understanding, it can be treacherous, holding us in the grip of false theories even in the face of robust recalcitrant data. Better to put genuine understanding on firmer (i.e. more "objective") ground, safe from phenomenological trickery.

But defending psychologism about understanding does not require equating genuine understanding with the sense of understanding. Rather, as Friedman argues, one can hypothesize that what is implicated in genuine understanding is one or more

[16] Newton-Smith writes:

> Dictionary definitions typically explicate the notion of explanation in terms or understanding: an explanation is something that gives understanding or renders something intelligible. Perhaps this is the unifying notion. The different types of explanation are all types of explanation in virtue of their power to give understanding. While certainly an explanation must be capable of giving an appropriately tutored person a psychological sense of understanding, this is not likely to be a fruitful way forward. For there is virtually no limit to what has been taken to give understanding. Once upon a time, many thought that the fact that there were seven virtues and seven orifices of the human head gave them an understanding of why there were (allegedly) only seven planets. We need to distinguish between real and spurious understanding. And for that we need a philosophical theory of explanation that will give us the hallmark of a good explanation. (Newton-Smith, 2001, p. 131)

psychological capacities that we all share. Waskan calls this capacity our capacity for making something "intelligible"—it is perfectly plausible to suppose that this capacity comes apart from the sense of understanding. Especially if we consider that whether a phenomenon is genuinely intelligible (in the sense described by Waskan) is a function not only of a particular psychological capacity, but on whether the deployment of this capacity is conducive to the ends of understanding. However, set aside for the moment what intelligibility might consist in. For it is not clear that this strategy addresses the original worry.

Recall that what motivates the subsistence argument where understanding is concerned is the supposed objectivity of explanatory knowledge—if understanding of the psychological variety varies from person to person, then this kind of understanding is not constitutive of explanatory knowledge. Now, grant that the *sense* of understanding varies from person. But note, the fact that we have a *capacity* for experiencing this sense, a capacity for finding something intelligible—is not something that varies from person to person. And the same is presumably true of any of our psychological capacities: while cognitive science targets cognitive capacities that we all share, it is surely the case that how well or to what extent those capacities are developed or put to use by different individuals varies significantly across those individuals. So, even supposing we can identify a capacity for intelligibility that we all share, precisely how that capacity functions will surely vary from person to person and case to case. So, the defender of anti-psychologism about understanding might argue, just as our capacity for experiencing a sense of understanding is not constitutive of genuine understanding—because the way it is deployed varies from person to person, it isn't any of our other psychological capacities that are what true understanding consists in either. This is because how these are deployed varies from person to person as well. If that is right, then Friedman's reply doesn't address the worry at the heart of the subsistence argument.

The defender of psychologism can, however, meet this objection by noting that on a psychologistic view, understanding is a cognitive state that is, under a certain range of conditions, conducive to certain ends—what determines whether a person has genuine understanding is whether the cognitive state the person has achieved actually is conducive to the achievement of those ends. On this view, understanding is objective in the sense that we share a certain cognitive capacity that *under the right conditions* is conducive to achieving the ends that are desirable where understanding is concerned. The philosophy and psychology of understanding are thus concerned with discovering the characteristic features of that state, its ends (or functions) *and* the conditions under which it satisfies that function. We have no guarantee that the capacity for understanding will succeed in achieving the ends that are its proper function—this is to be expected. Indeed, we should expect that individuals will differ (and widely) in the extent to which those states they have achieved are conducive to achieving the ends of understanding. This does not mean that psychologism about understanding is relative to the psychological features of various individuals. To the extent that we share certain cognitive capacities, these amount to constraints within which a system for solving certain problems must operate. Thus, given that understanding is properly conceived of in this way—as a means to achieving certain ends—the metaphysical objection to psychologism misses the mark.

5.3.2 The epistemological objection

Hempel's second argument against psychologism is epistemological: it challenges the view that psychological characteristics could figure in the standards for genuine understanding. Again, Hempel's concern is to challenge pragmatic accounts, according to which explanation requires that the individual for whom a description is an explanation also *understands* why the *explanandum* occurred, that is, finds the description familiar, illuminating, intelligible, and so on. According to Hempel, "a psychological factor of this kind certainly cannot serve as a standard in assessing the worth of a proposed explanation." (Hempel, 1965, p. 258) That is, the achievement of some psychological state cannot be what *justifies* the belief that explanatory knowledge, or genuine understanding, has been achieved. Hempel thus rejects, in Sober's terms, "epistemological psychologism" about understanding, i.e. the view that there is no question of justifying claims about understanding independently of what scientists actually seek as a matter of course in attempting to understand a phenomenon. He rejects such a view in favor of philosophically motivated norms for good explanations and scientific understanding, namely, that the explanandum (in the case of deductive nomological explanation) is a deductive consequence of initial conditions and the laws of nature.

Hempel provides two arguments for this claim. The first, as before, is that the standards for good explanations ought to be objective standards, but our psychological capacity for understanding "varies from person to person and from time to time."(*Ibid.*) As we have seen, this argument from the variation of our psychological attitudes falls short: the search for the psychological basis of understanding is a search for a psychological capacity that we share. And if psychologism succeeds then it does so by positing a capacity, one that does not vary "from person to person and from time to time", that figures into our justificatory practices. One is justified in believing that one has understanding when one has achieved a certain psychological state that conforms to well-established explanatory practices.

Hempel's second argument is, however, more compelling. This argument relies on the premise that the psychological state associated with understanding is neither necessary nor sufficient for genuine (scientific) understanding of the phenomenon to be explained. Therefore, such states could not figure in the standards for assessing claims about understanding. For example, as Hempel writes, in discussing scientific explanations of human action, specifically, "the view held by some scholars that the explanation, or the understanding, of human actions requires an empathetic understanding of the personalities of the agents":

> the existence of empathy on the part of the scientist is neither a necessary nor a sufficient condition for the explanation, or the scientific understanding, of any human action. (Hempel, 1965, pp. 257-8)

As noted above, the history of science is enough to show that the sense of understanding is not sufficient for understanding. And several examples serve to illustrate that a sense of understanding is not necessary for genuine understanding either.[17] In Chapter 4, I provided one such example in the Monty Hall problem. In this problem, most people's intuitions are exactly wrong about which move maximizes ones chances of winning. Even with the correct answer in hand, most people struggle to maintain a stable sense of understanding about the right answer to the problem. In another example, physicist Richard Feynman tells a lecture hall of students that they will never understand what he is going to teach them about quantum mechanics (Feynman, 1985)—not because they aren't smart enough, but because not even he understands quantum mechanics. Presumably, however, quantum mechanical theory is sufficient explanation for quantum phenomena. Finally, consider the so-called "explanatory gap" between physicalist conceptions of cognition and the nature of consciousness. One can imagine that a mature cognitive science (that includes philosophy) will ultimately provide an account of the nature, conditions, and causes

[17] Salmon raises basically the same objection:

> not only is there a danger that people will feel satisfied with scientifically defective explanations; there is also the risk that they will be unsatisfied with legitimate scientific explanations. A yearning for anthropomorphic explanations of all kinds of natural phenomena—for example, the demand that every explanation involve conscious purposes—sometimes leads people to conclude that physics doesn't *really* explain anything at all...Some people have rejected explanations furnished by general relativity on the ground that they cannot visualize a curved four-dimensional space-time. (Salmon, 1984, p. 13)

of consciousness, and yet one that leaves us without a feeling of satisfaction where explanation is concerned. Nevertheless, many philosophers have argued that the mature theory would be a genuine explanation of consciousness.[18]

However, as before, Hempel's argument equates the psychology of understanding with the *sense* of understanding. And given that the sense of understanding does not exhaust the candidate psychological features characteristic of understanding, then from the fact that "a psychological factor of this kind cannot serve as a standard" it does not follow that no psychological factor can. On the other hand, also as before, this opening to psychologism may be short-lived, for the question becomes whether Hempel's epistemological objection applies to psychologism more generally, and not merely its subjective phenomenological form. Even supposing that we can find an objective cognitive capacity that is implicated in understanding, there is no guarantee that this capacity is either sufficient or necessary for *genuine* understanding. That is, the argument goes, from the mere fact that our cognitive constitution is built in a certain way, we are not licensed to conclude that its proper functioning constitutes genuine understanding. Rather, on this view, what is required are philosophical norms for evaluating whether proper functioning of a cognitive capacity constitutes understanding in any particular case.

Craver's Epistemological Argument

Carl Craver provides a recent example of this objection that is targeted at what Craver calls "representational" accounts of explanation, according to which "explanations explain by subsuming a phenomenon under a general representation, prototype, or schema." (Craver, 2007, p. 28) Specifically, Craver targets Churchland's

[18] Cf. Flanagan (1992) on the distinction between satisfactory and satisfying. Also Churchland (1996). Nothing of metaphysical consequence follows from the fact (if it is a fact) that no sense of understanding will accompany a future complete theory of consciousness. Indeed, with an account of the psychology of understanding in hand, we might provide a theory for why certain satisfactory explanatory hypotheses do not also produce a sense of understanding.

parallel distributed processing (PDP) model of explanatory understanding described above. Craver argues that representational accounts are "too weak to serve as a guide to the norms that distinguish good explanations from bad and complete explanations from incomplete."(*Ibid.*) This argument can be understood as targeting any account that posits certain psychological capacities (e.g. in this case, representational capacities) as either necessary or sufficient for genuine understanding. I will consider Craver's objection in further detail as a way of assessing the epistemological argument against psychologism.

Recall that on Churchland's view, explanatory understanding is a species of recognition insofar as it amounts to "activation of a particular prototype vector in a well-trained network". According to Churchland, prototype activation is "dramatically ampliative": the recognition that a phenomenon is an instance of a prototype provides one with a huge amount of information that is not given by the occurrence of the phenomenon itself. But whether one *actually* gains such information, one might think, depends on whether the prototype that gets activated is an accurate representation of the world. And this is the point of departure for Craver's criticism. According to Craver, while the prototype activation model is "an intriguing hypothesis about the psychology of understanding and about how scientists represent the world to themselves and to one another", it is not a normatively adequate theory of scientific explanation.[19] Craver draws this conclusion by considering what he takes to be the norms of scientific explanation implicit in certain example cases in the neurosciences.

First, argues Craver, Churchland's model is not *sufficient* for scientific explanation because it cannot distinguish between genuine and non-explanatory descriptions. That is, we can imagine cases where our psychological capacity for understanding

[19] Supposing that only scientific explanations can produce genuine understanding (and I take it that Craver would endorse such a view), it follows that it is not a normatively adequate theory of genuine understanding either.

(e.g. prototype activation in a well-trained network) is functioning perfectly normally, but where we don't have genuine understanding. For example, Craver argues, the PDP model cannot distinguish between etiological and mere temporal sequences. In attempting to explain the action potential by the release of neurotransmitters, scientists initially discovered that a rise of intra-cellular sodium concentration precedes neurotransmitter release. But it turns out that it is the rise in membrane voltage, not the rise in sodium concentration, that *explains* neurotransmitter release (Craver, 2007, p. 31). It is not clear, Craver argues, how Churchland's model, which takes "etiological prototypes" to depict temporal associations between events, has the resources to make this distinction. Indeed, Churchland concedes that one would have to provide an account of how the brain distinguishes between etiological and mere temporal sequences. But as Craver writes:

> The way to understand how brains distinguish causes from temporal sequences is to start by considering how causes differ from temporal sequences—that is, by examining the objective explanations in the world rather than the way that they are represented in the mind/brain. (Craver, 2007, p. 31)

In other words, the psychology of understanding in this case is not relevant to what justifies claims about what explains neurotransmitter release. It is our knowledge of the causal features of the world.

Furthermore, the prototype-activation model does not, according to Craver, have the resources to distinguish between explanatorily relevant and irrelevant features of an explanation. Prototype vectors are representations of multiple dimensions of information of a system, but only some of this information is relevant to explaining a given phenomenon. In the above example, a rise in sodium concentration is part of a complete description of the processes that occur during neurotransmitter release.

But the rise in sodium concentration is not relevant to neurotransmitter release. So mere activation of a person's "detailed and well-informed representation" of the processes that occur during an action potential are not sufficient for identifying the explanatorily relevant features.

In both of these cases (distinguishing between explanatory and mere temporal sequences and identifying explanatorily relevant features of a system), the information required for assessing the explanation, and so the understanding that it produces, is, according to Craver, causal information. But what causal relations there are has nothing to do with human psychology. So merely providing an account of the psychology of understanding, an account of the proper functioning of a particular cognitive capacity, is not sufficient for scientific explanation, and so genuine understanding.

Furthermore, Craver argues, prototype activation is not *necessary* for scientific explanation. This is because sometimes a system is so complex that it overwhelms the brain's computational capacities so that the system cannot be sufficiently represented to capture the explanatorily relevant information. In such cases, subsuming a phenomenon under a representation becomes impossible. It remains the case, however, that there are features of the world, however complex, that caused the phenomenon to be explained, and so, on the causal view, these feature explain it; that is, these "explanations exist even if we cannot represent them cognitively." (Craver, 2007, p. 34) Of course, even if prototype activation isn't necessary for scientific explanation, it might still be necessary for genuine *understanding*. Any theory of understanding must take as at least a minimal commitment that some psychological capacity is necessary for understanding; after all, understanding is a psychological phenomenon. But even so, on Craver's view, the decisive justificatory requirement does not depend on the characteristic features of this capacity, but rather on the truth or accuracy of the representation that this capacity instantiates.

Reply to the Epistemological Objection

The reply to the epistemological objection to psychologism is, by now, a familiar one: it amounts to a defense of the naturalistic approach to theorizing about understanding. One thing that must be done to make this approach plausible is to distinguish epistemological psychologism (or, methodological naturalism) from the uncritical descriptivism with which it is often confused, that is, the view that theorizing about explanation (and understanding) amounts to describing the psychological capacities that human beings employ in coming to understand a phenomenon. But this is a mistake. As I have argued, epistemological psychologism is not an abdication of normativity. It is merely a rejection of its traditional form.

Consider Churchland again. In describing his aim in developing the prototype-activation model of explanatory understanding, he states that he "shall approach the topic with the aims of an empirical scientist rather than with the aims of a logician or conceptual analyst" and that his concerns are not primarily normative, but rather, descriptive in nature.[20] Craver seizes on this point, arguing that "the goal of thinking more clearly about what is required of an adequate neuroscientific explanation cannot be satisfied without thinking about norms for evaluating explanations." (Craver, 2007, p. 31) According to Craver, the vector prototype activation account does not satisfy this requirement so it is not an adequate philosophical theory of explanation

[20] Full quote:

> I shall approach the topic with the aims of an empirical scientist rather than with the aims of a logician or conceptual analyst. The goal is to outline a substantive empirical theory of what explanatory understanding really is, rather than to provide an analysis of the concept of explanation as it currently is used or ideally should be used. Normative issues will certainly be addressed, and some unexpected insights will emerge, but normative concerns are not the prime focus of the present chapter. Rather, what concerns us is the nature of the cognitive process that takes place inside the brain of the creature for whom explanatory understanding suddenly dawns, and in whom it is occasionally reactivated. (Churchland, 1989, p. 198)

and understanding.[21]

If Churchland really does endorse such a position, then Craver is right to criticize him for it, for in conceding normativity he concedes too much. But Churchland should not be read in this way. Recall especially his claim that the ampliative nature of explanatory understanding (the "welcome talent of ampliative recognition") is what makes it so *useful* for creatures such as ourselves. If the prototype-activation model is a plausible account of the psychology of understanding, then the standard for whether a given prototype activation constitutes genuine understanding is precisely whether it proves useful to the creature who instantiates it, that is, whether it achieves the ends of understanding. In other words, a more charitable reading of Churchland's aims, I suggest, is as providing one important component of a psychologistic account of explanatory understanding—namely, an account of the strategies that the brain adopts for solving the problem of achieving the ends of understanding. Along these lines, the following point deserves strong emphasis: epistemological psychologism does not endorse the view that the mere proper functioning of a cognitive capacity is sufficient for understanding. It rejects this view. Rather, genuine understanding requires that our capacity for understanding be conducive to the achievement of ends. And whether it is conducive to such ends will depend not only on the characteristic features of this capacity, but also on the world.

Suppose Churchland's model is a good one: we ought to expect that certain prototypes, when activated under a range of circumstances, will be well-established means of achieving those ends. To the extent that a particular explanatory context falls in that range, we are justified in believing that the activation of a particular prototype constitutes genuine understanding. On this view, prototype activation is sufficient for understanding just in case it constitutes a well-established means for achieving

[21] One can clearly see echoes of Salmon's criticism of psychological accounts of explanation as being "patently inadequate" to the normative demands of any plausible theory (Salmon, 1984).

the ends of understanding in the range of circumstances to which the explanatory context applies. Craver, and causal theorists more generally, are surely right that describing or otherwise representing the causes of a phenomenon—or in Churchland's language, the activation of etiological prototypes—is, under a wide range of circumstances, a well-established means of achieving the ends of understanding. But on the view defended here, they are not the only well-established means of achieving those ends. Furthermore, they are effective within a particular range of circumstances, and outside of this range they are less effective.[22]

If the prototype-activation model of explanatory understanding is a plausible account of the psychology of understanding, is prototype-activation necessary for understanding? Before answering this question, consider the multi-layered nature of Churchland's model. From the neurophysiological or implementation perspective, Churchland provides a broadly connectionist account of how the brain is able to represent and exploit information about the external world. In particular, Churchland argues that this kind of system is a highly effective means for large populations of neurons of executing complex functions at a very high speed—it exploits the ability of neurons to make thousands of dendritic connections to the axon terminals of other neurons to instantiate a parallel processing system of incredible computational and representational power. Churchland describes this power in vivid terms: a conservative estimate of the number of distinct activation vectors that any given functionally specialized subsystem of the brain can instantiate is on the order of $10^{100,000,000}$. As Churchland writes, "To appreciate the magnitude of this number, recall that the total number of elementary particles in the entire physical universe, photons included, is only about 10^{87}." (Churchland, 1989, p. 209)

If Churchland is right that the brain is well-described in terms of a prototype-

[22] We should expect to know exactly what that range is—part of doing science is exploring the boundaries of applicability of various explanatory strategies.

activation model, then it is somewhat trivial to note that instantiating a connectionist prototype activation system is necessary for understanding. According to Churchland, that's just what the brain is. In that sense, the brain represents a particular kind of design solution to the problem of achieving certain ends within the design constraints imposed by our evolutionary history. Furthermore, it appears to be a very effective means of achieving those ends under a wide range of circumstances. One might wonder whether there are other potential design solutions to the problem—for example, we might hypothesize about alternative methods of representing and computing that are effective at achieving our ends. Or one might wonder under what circumstances the brain falls short of these aims. These are both interesting questions. But to the extent that we are interested in what it means for human beings to understand something, the building blocks of human cognition will be a prerequisite for the cognitive capacities we have.[23]

From the perspective of the cognitive capacities that actually get implemented, the question becomes whether the prototype kinds themselves are necessary for understanding. In other words, the question is whether the activation of a particular prototype is the best means, given the constraints of the system, of achieving the ends of understanding. But even here, this starts to look like the wrong kind of question. Different prototypes will be effective at achieving ends under different ranges of circumstances. But even within a particular range, it would be difficult to conclude that we have hit upon the very best strategy for achieving our ends in those circumstances. What vector prototypes are instantiated by the system is not set in stone and it is open to us to invent new ways of representing a system, ways that might be more effective under certain ranges of circumstances. When we hit upon

[23] Perhaps the point is best made by distinguishing between implementation and what gets implemented. In this case, a connectionist system implements a prototype activation process. What we are really interested in seems to be whether the activation of certain prototypes is necessary for understanding.

a strategy that is effective at achieving certain ends, we are justified in thinking we have achieved understanding. But we shouldn't think that this particular strategy is the only strategy available to us, and so we shouldn't think that it is, strictly speaking, necessary for understanding.[24]

Consider this last point in light of Craver's criticism that explanatory understanding is not necessary for scientific explanation—in particular, because a system might be so complex that it cannot be represented cognitively. On Craver's view, explanation in cognitive neuroscience is mechanistic. A consequence of this view is that those mechanisms that are so complex that they can't be adequately represented can't be understood. And yet they are still what explains the phenomenon of interest. On the view I am defending here, explanation and understanding are not like this. Rather, explanation and understanding are intertwined: they are ways of describing a system (the brain) that is at once both a design solution in itself as well as an active designer of solutions, solutions to achieving the ends of understanding. On this view, explanations are strategies for achieving understanding, where understanding is a cognitive capacity that is conducive to achieving certain ends under ranges of circumstances—this is what unifies these strategies under the heading "explanation". So, on this view, there is no explanation without at least the possibility of understanding.

It remains possible that, due to our cognitive limitations, there are some systems that are cognitively intractable where understanding is concerned. But this would not be a mere matter of complexity. Indeed, we appear to have methods of representing complex systems that are effective means of achieving our understanding ends. Rather, it would mean that no cognitive representation of the system in question

[24] Note: thinking of this model in terms of different layers I think allows us to focus on a more cognitivist approach to understanding that is in line with how cognitive neuroscience actually seems to proceed. We should use our understanding of neurophysiology (i.e. of mechanisms) together with insights from the full range of sub-disciplines in the cognitive sciences to develop models of our cognitive capacity for understanding.

would allow us to achieve those ends. Of course, it is difficult to imagine what such a scenario would look like. Perhaps there are certain features of quantum mechanics or the science of consciousness that fit this scenario and our lack of a sense of understanding about certain aspects of these cases is an indicator of falling short in some way. But note, it is always open to us in such cases to say that we have just not imagined the right way of thinking about these problems—we do not have an appropriate vector prototype, or model of the system. And a current failure of imagination would not appear to entail any lasting metaphysical conclusions about our own cognitive limitations.

My aim here is not to defend Churchland's neuropsychological account of understanding in particular—I discussed this case because Craver targets the view specifically. It is rather to defend the more general view that justifying claims about understanding involves identifying cognitive strategies for understanding, that is, those strategies that are conducive to achieving the ends of understanding under a particular range of circumstances and within our psycho-biological constraints. Churchland's own account is, I take it, an example of a particular neurocognitive strategy for doing just this. It is further question, of course, whether Churchland is right that the mind-brain is well-described by this model, whether this model has, as Sober would say, psychological reality.

6

Conclusion

The guiding question of this dissertation is the question how cognitive neuroscience explains. In attempting to shed light on this question, I criticized a prominent view of explanation in the mind sciences, according to which explaining a cognitive phenomenon consists in discovering and describing the mechanisms responsible for the phenomenon under investigation. I further argued that the failure of the mechanist program is symptomatic of the implausibility of the dominant theory of scientific understanding, according to which scientific understanding amounts to having knowledge of certain objective explanatory relations. Supposing I have been successful in these critical aims, how then does cognitive neuroscience explain? I have argued that it explains by constructing models that amount to effective strategies, understood relative to the psycho-biological constraints of human cognition, for achieving a variety of ends, ends that we value when we seek understanding. I shall conclude with a brief summary of possible future research that builds on the work completed here.

6.1 Scientific Understanding

In Chapter 4, I outlined a normative framework for investigating the nature of scientific understanding, conceived as set of cognitive capacities that, under the right conditions, are effective strategies for achieving certain ends. This framework calls for an interdisciplinary and empirically-informed approach to addressing three central questions regarding the nature of understanding: What are the ends of understanding? What is the nature of those cognitive capacities for understanding such that they are conducive to achieving these ends? And under what conditions are these capacities an effective means of achieving these ends—i.e. when is the deployment those capacities justified? One way to shed light on these questions will be to investigate additional areas of research in cognitive neuroscience, for example, research on perception, language, and social cognition.

6.2 Cognitive Models

A model-based approach to the nature of understanding is a promising approach to fulfilling the investigative strategy described above. This approach has a well-established tradition in the philosophy of science as an alternative to the view that understanding is a traditional kind of propositional knowledge and it is an active area of research in the philosophy of science. However, further research on the nature and function of models is needed along two dimensions. First, models can be characterized as abstract objects and research on their pragmatic, inferential, and semantic properties is needed. For example, how should we understand the relationship between model and world? Cartwright and Weisberg argue that similarity is the appropriate relational criterion. An important question becomes whether this appeal to similarity is appropriate to the demands of the account of understanding I have defended. Second, whether abstract objects of this kind play an actual role in human cognition, whether the mind-brain instantiates them and if so, how it recruits them toward achieving the ends of understanding, are very interesting questions. The question what constitutes the ends of understanding and the question how a model is conducive to these ends go hand-in-hand: we can ask what features models possess such that they are conducive to certain hypothesized ends; and we can ask which are those ends that the characteristic features of models are conducive to achieving. Evaluating and developing the model-based approach will require integrating research on the nature of models as abstract objects with research in a range of empirical disciplines.

6.3 The Nature of Mind and Cognition

Another important questions concerns how the account of explanation and understanding in the mind sciences that I have defended bears on philosophical questions about the nature of mind and cognition. For example, intentional notions akin to beliefs and desires are central to the neuroeconomic model of decision-making. How should appeal to such notions be understood from the perspective of the framework for investigating understanding that I have proposed? For example, it seems that on this account, the explanatory framework of reasons could be incorporated into a unified account of explanation and understanding without requiring that it be assimilated by the explanatory strategies characteristic of, for example, physics or chemistry. Furthermore, to the extent that the framework of reasons incorporates information about our phylogenetic and ontogenetic histories, we have an account of why it has proven indispensable in both ordinary discourse and in our best cognitive science. Investigating these issues will require broadening this research into the philosophy of action and the social sciences.

Bibliography

Batterman, R. (2002). Asymptotics and the role of minimal models. *British Journal for the Philosophy of Science*, 53(1):21–38.

Batterman, R. (2010). On the explanatory role of mathematics in empirical science. *British Journal for the Philosophy of Science*, 61(1):1–25.

Batterman, R. W. (2009). Idealization and modeling. *Synthese*, 169(3):427–446.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. Psychology Press. Perkins QP360.5 .B43 2008.

Bechtel, W. and Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2):421–441.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1):pp. 23–36.

Blair, R. J. R. (2001). Neurocognitive models of aggression, the antisocial personality disorders, and psychopathy. *Journal of Neurology, Neurosurgery & Psychiatry*, 71(6):727–731.

Bromberger, S. (1966). Questions. *Journal of Philosophy*, 63(20):597–606.

Brown, H. I. (1988). *Rationality*, volume 100. Routledge.

Carnap, R. (1974). *An Introduction to the Philosophy of Science*. Dover.

Cartwright, N. (1983). *How the laws of physics lie*. Cambridge Univ Press.

Cartwright, N. (1989). *Natures capacities and their measurement*. Clarendon.

Chemero, A. and Silberstein, M. (2008). After the philosophy of mind: Replacing scholasticism with science. *Philosophy of Science*, 75(1):1–27.

Churchland, P. M. (1989). On the nature of explanation: A pdp approach. In *A Neurocomputational Perspective*. MIT Press.

Churchland, P. S. (1996). The hornswoggle problem. *Journal of Consciousness Studies*, 3(5-6):402–8.

Cohen, I. B. (1978). *Introduction to Newton's 'Principia'*. Harvard University Press.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, 4(3):317–370.

Craver, C. F. (2003). The making of a memory mechanism. *Journal of the History of Biology,*, 36(1):153–195.

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

Craver, C. F. (2007). *Explaining the brain : mechanisms and the mosaic unity of neuroscience*. Oxford : Clarendon Press ; New York : Oxford University Press.

Craver, C. F. and Kaplan, D. M. (2011). Towards a mechanistic philosophy of neuroscience. In *Continuum Companion to the Philosophy of Science*, page 268. Continuum.

Cummins, R. C. (1975). Functional analysis. *Journal of Philosophy*, 72(November):741–64.

Cummins, R. C. (2000). "how does it work" versus "what are the laws?": Two conceptions of psychological explanation. In Keil, F. and Wilson, R. A., editors, *Explanation and Cognition, 117-145*. MIT Press.

Davidson, D. (1973). On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association*, 47(n/a):5–20.

de Regt, H. (2009). Intelligibility and scientific understanding. In Regt, H. D., Leonelli, S., and Eigner, K., editors, *Scientific Understanding: Philosophical Perspectives*. University of Pittsburgh Press.

Dennett, D. C. (1978). *Brainstorms*. Number 121. MIT Press.

Dennett, D. C. (1991). Real patterns. *Journal of Philosophy*, 88(1):27–51.

Dray, W. H. (1979). *Laws and Explanation in History*, volume 10. Greenwood Press.

Dretske, F. (1989). Reasons and causes. *Philosophical Perspectives*, 3:1–15.

Dretske, F. I. (1988). *Explaining behavior : reasons in a world of causes*. Cambridge, Mass. : MIT Press, c1988.

Duhem, P. M. M. (1991). *The Aim and Structure of Physical Theory*, volume 13. Princeton University Press.

Feynman, R. P. (1985). *QED. The strange theory of light and matter*, volume 1. Princeton University Press, Princeton, NJ, 7th printing, with corrections, 1988 edition.

Flanagan, O. (1992). *Consciousness Reconsidered*. Bradford Books. MIT Press.

Flanagan, O. (2006). Varieties of naturalism. In Simpson, P. C. . Z., editor, *The Oxford Handbook of Religion and Science*, pages 430–452. Oxford University Press.

Fodor, J. A. (1991). The dogma that didn't bark (a fragment of a naturalized epistemology). *Mind*, 100(2):201–220.

Friedman, M. (1953). *Essays in positive economics*. University of Chicago Press.

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, 71(1):5–19.

Garfinkel, A. (1981). *Forms of explanation : rethinking the questions in social theory*. New Haven, Conn. : Yale University Press.

Giere, R. N. (2010). *Explaining science: A cognitive approach*. University of Chicago Press.

Glimcher, P. W. (2011). *Foundations of neuroeconomic analysis*. Oxford University Press Oxford.

Goodman, N. (1983). *Fact, Fiction, and Forecast*. Harvard University Press.

Gopnik, A. (1998). Explanation as orgasm. *Minds and machines*, 8(1):101–118.

Hacking, I. (1980). Strange expectations. *Philosophy of Science*, 47(4):562–567.

Hempel, C. G. (1965). Aspects of scientific explanation and other essays in the philosophy of science. *New York*.

Hitchcock, C. R. (1992). Causal explanation and scientific realism. *Erkenntnis*, 37(2):151–178.

Iyengar, S. S. and Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology*, 79(6):995.

James, W. (2001). *Psychology: The Briefer Course*. Dover Publications.

Jones, M. R. (2005). Idealization and abstraction: A framework. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, 86(1):173–218.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1990). Experimental tests of the endowment effect and the coase theorem. *Journal of political Economy*, pages 1325–1348.

Kahneman, D. and Tversky, A. (1984). Choices, values, and frames. *American psychologist*, 39(4):341.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, 183(3):339–373.

Kaplan, D. M. and Bechtel, W. (2011). Dynamical models: An alternative or complement to mechanistic explanations? *Topics in Cognitive Science*, 3(2):438–444.

Kaplan, D. M. and Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, 78(4):601–627.

Khalifa, K. (2012). Inaugurating understanding or repackaging explanation? *Philosophy of Science*, 79(1):15–37.

Kim, J. (1988). What is "naturalized epistemology?". *Philosophical Perspectives*, 2:381–405.

Kuhn, T. S. (1996). *The structure of scientific revolutions.* Chicago, IL : University of Chicago Press, 1996.

Kyburg, H. E. (1965). Salmon's paper. *Philosophy of Science*, 32(2):147–151.

Lakatos, I. (1971). History of science and its rational reconstructions. In Buck, R. C. and Cohen, R. S., editors, *Psa 1970. Boston Studies in the Philosophy of Science Viii*, volume 1970, pages 91–108. D. Reidel.

Lange, M. (2013). What makes a scientific explanation distinctively mathematical? *British Journal for the Philosophy of Science*, 64(3):485–511.

Levy, A. and Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science*, 80(2):241–261.

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25.

Martin, R. (2013). The st. petersburg paradox. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2013 edition.

McFadden, D. L. (2005). Revealed stochastic preference: a synthesis. *Economic Theory*, 26(2):245–264.

Mill, J. S. (1956). *A system of logic: Ratiocinative and inductive.* Longmans, Green and Company.

Milner, D. and Goodale, M. (1998). Prcis of the visual brain in action. *Psyche*, 4 (12):1–14.

Montague, P. R., Hyman, S. E., and Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature*, 431(7010):760–767.

Neander, K. (1991). Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science*, 58(2):168–184.

Nersessian, N. J. (1984). Aether/or: The creation of scientific concepts. *Studies in History and Philosophy of Science Part A*, 15(3):175–212.

Nersessian, N. J. (1995). Constructive modeling in creating scientific understanding. *Science and Education*, 4:203–226.

Nersessian, N. J. (2006). Model-based reasoning in distributed cognitive systems. *Philosophy of Science*, 73(5):699–709.

Newton, I. (1999). *The Principia : Mathematical Principles of Natural Philosophy.* University of California Press.

Newton-Smith, W. (2001). Explanation. In Newton-Smith, W., editor, *A Companion to the Philosophy of Science*, Blackwell Companions to Philosophy, chapter 19, pages 127–133. Wiley.

Norman, K. A. and O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological review*, 110(4):611.

Peirce, C. S. (1966). *Selected Writings Edited, with an Introd. And Notes.* Dover Publications.

Piccinini, G. and Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183(3):283–311.

Platt, M. L. and Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, 400(6741):233–238.

Purves, D. and Lotto, R. B. (2003). *Why we see what we do: An empirical theory of vision.* Sinauer Associates.

Quine, W. (1973). *Word and object.* The MIT Press Paperback Series. The M.I.T. Press.

Quine, W. V. (1969). Epistemology naturalized. In *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Quine, W. V. and Ullian, J. (1970). *The Web of Belief*. Random House.

Railton, P. (1978). A deductive-nomological model of probabilistic explanation. *Philosophy of Science*, pages 206–226.

Railton, P. (1981). Probability, explanation, and information. *Synthese*, 48(2):233–256.

Reber, A. S. (1993). Implicit learning and tacit knowledge: An essay on the cognitive unconscious. *Oxford University Press*.

Rice, C. (2013). Moving beyond causes: Optimality models and scientific explanation. *Nous*, 48(2).

Rosenberg, A. (1996). A field guide to recent species of naturalism. *British Journal for the Philosophy of Science*, 47(1):1–29.

Ruben, D.-H. (1990). *Explaining Explanation*. Routledge.

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

Salmon, W. C. (1998). *Causality and Explanation*. Oxford University Press.

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, pages 61–71.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):pp. 1593–1599.

Scriven, M. (1962). Explanations, predictions, and laws. *Minnesota studies in the philosophy of science*, 3:170–230.

Selten, R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games. *International journal of game theory*, 4(1):25–55.

Shadlen, M. N. and Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proceedings of the National Academy of Sciences*, 93(2):628–633.

Simon, H. A. (1957). Models of man; social and rational.

Sinnott-Armstrong, W. and Fogelin, R. (2015). *Understanding Arguments: An Introduction to Informal Logic*. Cengage Learning.

Smith, D. V. and Huettel, S. A. (2010). Decision neuroscience: neuroeconomics. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):854–871.

Sober, E. (1978). Psychologism. *Journal for the Theory of Social Behaviour*, 8(July):165–91.

Stich, S. P. (1990). *The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation.* The MIT Press.

Strevens, M. (2008). *Depth : an account of scientific explanation*. Cambridge, Mass. : Harvard University Press.

Thagard, P. and Nisbett, R. E. (1983). Rationality and charity. *Philosophy of Science*, 50(2):250–267.

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, 69(2):212–233.

Trout, J. D. (2007). The psychology of scientific explanation. *Philosophy Compass*, 2(3):564–591.

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

van Gelder, T. (1995). What might cognition be if not computation? *Journal of Philosophy*, 92(7):345–81.

Von Eckardt, B. (1995). *What is cognitive science?* MIT press.

Von Eckardt, B. and Poland, J. S. (2004). Mechanism and explanation in cognitive neuroscience. *Philosophy of science*, 71(5):972–984.

von Neumann J. and O., M. (1944). *Theory of games and economic behavior*. Princeton University Press, Princeton.

Waskan, J. (2008). Knowledge of counterfactual interventions through cognitive models of mechanisms. *International Studies in the Philosophy of Science*, 22(3):259–275.

Waskan, J. (2011). Intelligibility and the cape: Combatting anti-psychologism about explanation. In *Epistemology of Modeling & Simulation: Building Research Bridges between the Philosophical and Modeling Communities.* Pittsburgh.

Weisberg, M. (2006). Forty years of 'the strategy': Levins on model building and idealization. *Biology and Philosophy*, 21(5):623–645.

Weisberg, M. (2007a). Three kinds of idealization. *Journal of Philosophy*, 104(12):639–659.

Weisberg, M. (2007b). Who is a modeler? *British Journal for the Philosophy of Science*, 58(2):207–233.

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

Woodward, J. (2003). *Making things happen : a theory of causal explanation*. New York : Oxford University Press, 2003.

Woodward, J. (2011). Scientific explanation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2011 edition.

Woodward, J. and Hitchcock, C. (2003). Explanatory generalizations, part i: A counterfactual account. *Noûs*, 37(1):1–24.

Wright, L. (1976). *Teleological Explanations: An Etiological Analysis of Goals and Functions*. University of California Press.

Biography

Gordon J. Steenbergen was born on August 11, 1974 in Silver Spring, MD where he lived until graduating from high school, moving to Ann Arbor, MI to study engineering. He received a Bachelor of Science in Electrical Engineering from the University of Michigan. After a year in industry, he moved to England to study philosophy at the University of Birmingham, where he received a Postgraduate Diploma in Philosophy. Gordon spent the next several years alternating between graduate work in philosophy and working as a communications engineer, mostly as a contractor for the U.S. Army. Gordon earned his Master of Arts in Philosophy from Tufts University and his Ph.D. in Philosophy from Duke University. He lives in Guatemala City with his wife Becky and their three daughters.