Explaining and Understanding

DISSERTATION

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

By

Daniel Akiva Wilkenfeld

Graduate Program in Philosophy

The Ohio State University

2013

Dissertation Committee:

Professor Robert Kraut (Co-Advisor)

Professor Stewart Shapiro (Co-Advisor)

Professor Richard Samuels

Associate Professor Declan Smithies
Abstract

I argue for a new approach to the philosophy of explanation, one that treats explaining as an activity defined functionally by its propensity to produce understanding. Since this account is best articulated along with an account of understanding, I develop one. Specifically, I argue that understanding is best construed as a way of representing what is understood in a way that allows one to make efficacious inferences and successfully manipulate objects in the world. This pair of views stands in contrast with traditional views in the philosophy of science, which have been largely concerned with characterizing explanation in terms of some favored structure (e.g. deductive arguments) or content (e.g. reference to causes or mechanisms). I also explore the relationship between understanding and knowledge, arguing that understanding does not depend on justified belief. Finally, I address the recalcitrant mystery of explanatory asymmetry, arguing that later facts can sometimes explain earlier ones.
Dedication

To AJ. Simply. Phenomenal.
Acknowledgments

I would like to thank The Ohio State University for its support over the last seven years. In addition to my committee members, all faculty members of the philosophy department provided invaluable feedback. I would also thank my family for their support, in particular my wife Rachael, without whom this would not have been possible.
Vita

2002......................................................Fuchs Mizrachi School

2006......................................................B.A. Philosophy, Amherst College

2009......................................................M.A. Philosophy, The Ohio State University

2006 to present ......................................Graduate Teaching Associate and Graduate

   Fellow, Department of Philosophy, The

   Ohio State University

Publications

“Understanding as Representation Manipulability,” Synthese 190.6 (April 2013): 997-
   1016.

Fields of Study

Major Field: Philosophy
Table of Contents

Abstract ........................................................................................................................................ ii
Dedication ..................................................................................................................................... iii
Acknowledgments ......................................................................................................................... iv
Vita................................................................................................................................................ v
Publications .................................................................................................................................... v
Fields of Study .............................................................................................................................. v
Table of Contents ........................................................................................................................ vi
List of Tables ................................................................................................................................... x
Chapter 1: Why All This? ............................................................................................................. 1

Chapter 2: A (Biased) History of the Philosophy of Explanation .............................................. 7

§1 The Deductive Nomological Account .................................................................................... 8

§2 A Statistical Detour—The Quantum Mechanical Threat .................................................... 18

§2.1 Salvaging D-N from Electrons ............................................................................................. 18

§2.2 Statistical-Relevance Explanation: Drifting Towards Causation ...................................... 22

§3 Causation Proper ..................................................................................................................... 29

§4 Unificationist Accounts: Drifting Towards Functional Characterization .............................. 35
| §5 Pragmatic Accounts: Drifting Towards Understanding | 38 |
| §6 Conclusion: Coming Ashore | 41 |

Chapter 3: Functional Explaining: A New Approach to the Philosophy of Explanation | 43 |
| §1 Explanation, Explaining, and Understanding | 46 |
| §2 Hempel & Oppenheim's Enduring Omission | 53 |
| §3 The Respectability of Understanding | 55 |
| §4 Functional Explaining and the Philosophy of Scientific Explanation | 57 |
| §4.1 D-N & Unificationism | 60 |
| §4.2 Causal Accounts | 64 |
| §4.3 Pragmatic Accounts | 67 |
| §4.4 A Response to an Objection | 72 |
| §5 Understanding | 73 |
| §6 Conclusion | 77 |

Chapter 4: A Review of Understanding | 78 |
<p>| §1 Understanding and Explanation | 81 |
| §2 Understanding and (Explicit) Knowledge | 87 |
| §2.1 Understanding as Knowledge of an Answer | 88 |
| §2.2 Understanding as a Global Property of a System of Knowledge | 93 |
| §3 Understanding and Ability | 95 |</p>
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>§4 The Phenomenology of Understanding</td>
<td>102</td>
</tr>
<tr>
<td>§5 Understanding and Representation</td>
<td>107</td>
</tr>
<tr>
<td>§6 No Understanding?</td>
<td>107</td>
</tr>
<tr>
<td>Chapter 5: Understanding as Representation Manipulability</td>
<td>111</td>
</tr>
<tr>
<td>§1 Some Preliminary Characteristic Sentences</td>
<td>112</td>
</tr>
<tr>
<td>§2 Understanding Within and Without Philosophy of Science</td>
<td>114</td>
</tr>
<tr>
<td>§3 The Positive Proposal</td>
<td>120</td>
</tr>
<tr>
<td>§3.1 The Account</td>
<td>121</td>
</tr>
<tr>
<td>§3.2 Context-Sensitivity</td>
<td>129</td>
</tr>
<tr>
<td>§4 URM in the Field</td>
<td>134</td>
</tr>
<tr>
<td>§4.1 Understanding in Psychology</td>
<td>134</td>
</tr>
<tr>
<td>§4.2 Understanding in Artificial Intelligence</td>
<td>137</td>
</tr>
<tr>
<td>§5 Objections and Replies</td>
<td>138</td>
</tr>
<tr>
<td>§6 Conclusion</td>
<td>143</td>
</tr>
<tr>
<td>Chapter 6: Understanding sans Justification</td>
<td>145</td>
</tr>
<tr>
<td>§1 What Understanding? What Knowledge?</td>
<td>146</td>
</tr>
<tr>
<td>§2 Justification—a Needlessly High Bar</td>
<td>149</td>
</tr>
<tr>
<td>§2.1 Understanding, Justification, and the Challenger Explosion</td>
<td>149</td>
</tr>
<tr>
<td>§2.2 Theory-Driven Reasons to Treat Understanding before Justification</td>
<td>153</td>
</tr>
</tbody>
</table>
§3 (2) or (3)? ................................................................. 158

§3.1 Is Understanding a Species of Knowledge? ........................................... 159

§3.2 Knowledge without Justification .......................................................... 164

§4 Conclusion ............................................................................................. 168

Chapter 7: Explanatory Asymmetry and Pragmatic Explanation ..................... 170

§1 Explanatory Asymmetries ........................................................................ 171

§2 Van Fraassen’s Tower .............................................................................. 175

§3 A Simple Example .................................................................................. 176

§4 Causal Accounts and Explanatory Asymmetries ........................................ 180

Chapter 8: Taking FE Out for a Spin .............................................................. 183

§1 Some Surprising Results Regarding Theism .............................................. 183

§2 What now? ............................................................................................. 185

References .................................................................................................. 188
List of Tables

Table 1 Not Worrisome Partition of Graduate School Applications .............................. 27
Table 2 Still Worrisome Partition of Graduate School Applications .............................. 28
Table 3 Explanation and Understanding--The Logical Space ............................................. 87
Chapter 1: Why All This?

People go around explaining facts, both specific and general. We explain why we were late for a meeting, why water freezes at 32 degrees Fahrenheit at sea level, why human beings possess a seemingly useless kidney, and how ethical facts seem both wholly objective and intrinsically motivating at the same time. We explain events: the attack on Benghazi, or the Patriots winning the Super Bowl. We sometimes explain objects: we explain great works of art or historical figures. For all that, it is surprisingly difficult to say what exactly we are doing when we explain anything.

At first glance, explaining bears a certain resemblance to merely informing a given audience about facts pertaining to what we are trying to explain. I explain why I am late by saying that my car unexpectedly broke, the properties of water by citing molecular mechanics, and kidney possession by describing evolutionary pressures. Not just any old facts about me, or perhaps even any old facts about the causal history leading up to my lateness, help to explain why I was late. I might very well have put on a red shirt this morning, yet in pointing that out I do not explain why I was late. Even pointing out that I dithered over my shirt choice for three hours might fail to explain why I was late, if the audience happens to know that I always get up six hours before meetings. Conversely, explanations sometimes do not seem to require presenting unknown facts at all. Even if
someone is aware of all relevant facts pertaining to substances undergoing phase changes at certain temperatures, we can still explain the dynamics by comparing the molecules to objects in a large-scale model.

Despite, or perhaps because of, the ubiquity of explanations, one might question whether we really need a theory of what explanations are—if we can give and evaluate them in everyday life, is there a need for a more abstract model of explanatory practice? There would, however, be several advantages to possessing such a model. One advantage to possessing a theory of explanation is that our general practice of identifying explanations, while sufficient in common paradigm cases, nevertheless can break down when applied to more borderline instances. For example, theists and atheists have been known to debate whether the existence of an all-powerful deity would explain the properties and very existence of the universe. A full theory of explanation might be expected to answer this question.¹

A second advantage of having an account of what explaining is is that it might help to improve our explanatory practice. Knowing what farm produce is made of is of course not necessary to growing better samples, but knowing the chemical makeup of corn might enable one to fertilize soil more efficiently. It is impossible to say in advance of knowing the details of an account of what x's are whether such an account will help in the production of x's. Similarly, an account of what explanations are might or might not prove valuable to the extremely important task of evaluating explanations (my own

¹ Sneak preview: it will vary from context to context, which fact might itself explain why theists and atheists seem unable to agree over so much as the very rules of the debate. See Chapter 8.
account, developed in Chapter 3, does extend in natural ways to an account of explanation evaluation). Again, whether or not it is so helpful will depend on the details.

The most important thing to know when evaluating explanations (or anything else) is what they are supposed to do, which might or might not prove related to what they fundamentally are. On my account of explanations, whether something counts as an explanation is largely a function of what it can be used for, so explanation evaluation will hew very closely to explanation identification.

There is also an old but rarely cited practical rationale for studying the philosophy of explanation and understanding—learning things is, generally speaking, enjoyable. David Hume cited this fact in defense of philosophy as a whole in the first chapter of his *Enquiry Concerning Human Understanding* (1739/2001, p. 91), but it bears particular relevance to the subject matter of this essay. Learning is, generally, enjoyable. Yet (I would conjecture that) it is not just any learning that contributes to human well-being; though I have no evidence to support this claim, I imagine that rote memorization—even of something as intellectual as the periodic table of elements—contributes minimally to the learner's happiness. In general, it seems, the best way to characterize those instances of learning that make us happy is as those occasions where we learn an explanation and/or gain understanding of a subject matter of interest. If that is correct, then the philosophy of explanation and understanding actually does double duty in the creation of human happiness. An essay such as this contains many putative explanations, and those that are successful will hopefully engender some understanding—thus, if the subject matter happens to be of interest, a well-done work on explanation and understanding can
create happiness directly. More generally, however, it will grant those interested in other matters foreknowledge of what sorts of learning are likely to be happiness-producing.

The final reason to produce a new work on explanation and understanding again traces to Hume's defense of philosophy generally (p. 92)—if one is of the opinion that others have reached incorrect conclusions in their philosophizing then, whatever the merits of the project originally, it is incumbent on one to try to do things better. There has been a lot of philosophizing regarding the nature of explanation, particularly within the last 65 years. In the next chapter I will look at the most well-known products of this exploration, and argue that all of them face significant challenges.

In Chapter 3 I present my own positive view (FE)—that explanations are (caveats aside) just those things that can be used to create understanding. Most other accounts are wrong not merely in their technical details, but in their general approach to explanation. Extant accounts have, by-and-large, attempted to identify explanations by some tell-tale structure or content when, I argue, we can get a much better view of them by looking at the role they play vis-a-vis understanding.

If the FE is correct, its relative simplicity might make it surprising that it has been for so long overlooked. This surprise is lessened when one notes that understanding\(^2\)—which in FE plays a critical role in defining explanation—is itself at least as mysterious a concept. Not only is the nature of understanding unclear, but it has been much overlooked in comparison to the question of the nature of explanation. The literature there is on

\(^2\) I use boldface to denote concepts.
understanding will be examined in Chapter 4 (though the relation between understanding and knowledge will hold till Chapter 6). All accounts proposed will be found wanting.

This leads naturally to the construction of a new account of understanding (understanding as representation manipulability—URM) in Chapter 5. This is where the inquiry tentatively bottoms out—while the notions URM uses to explicate understanding—representation\(^3\), efficacious inference, contextual relevance, object manipulation, and counterfactuals—are each in their own way fairly vexed, they are all well-worn and traditional problems. Thus, understanding will prove a well-grounded and scientifically/philosophically tractable concept, which in turn can be used to elucidate explanation.

In chapter 6 I consider the interrelations between understanding and theories about the nature of knowledge. The conclusion is that understanding does not obtain in virtue of possessing justified belief. Since knowledge, on most traditional accounts, does obtain in virtue of possessing justified beliefs, we get the potentially surprising result that either understanding is not a species of knowledge or knowledge does not obtain in virtue of possessing justified belief. I give some tentative reasons for favoring this latter position.

In chapter 7, I discuss in greater detail the apparent asymmetries (causal and temporal—see next chapter for examples) in explanation, and argue that they are best viewed as mere pragmatic conventions applicable in most (but not all) contexts.

\(^3\) Representation might be thought to be particularly problematic. I do not undertake a full defense of representations here, but in Chapter 8 I do attempt to ameliorate some philosophical concerns that I am relying on a defunct notion.
I conclude in Chapter 8 with a look at where the research trajectory implemented in this dissertation is most naturally taken.
Chapter 2: A (Biased) History of the Philosophy of Explanation

Since Hempel & Oppenheim's 1948 paper "The Logic of Scientific Explanation", philosophers have been interested in articulating a precise account of what explanation is. During this time at least a dozen specific accounts have been developed, focusing variously on the logical form of explanations, the real-world relationships presented in the content of explanations, and the pragmatic context of giving and receiving explanations.

It has been accepted from Hempel & Oppenheim onward that an account of explanation should ideally provide necessary and sufficient conditions for something being an explanation. In attempting to give such conditions, the accounts became open to being discredited by what Jaegwon Kim (1994, p. 52) calls "the method of counterexamples": we possess pre-theoretic intuitions that certain paradigms count as explanations whereas other things (paradigmatically) do not. If a consequence of an account is that a paradigmatic explanation does not count as an explanation or a paradigmatic non-explanation does, the account is taken to be refuted. Remarkably, the vast majority of extant accounts of explanation face abundant counterexamples of both types. In the next chapter I suggest why this is the case and what the solution should be. (As a preview, the claim in the next chapter is that almost every extant account falls prey to counterexamples for the same reason, which is that the conceptual tools utilized to
construct the account have been uniformly deficient.) In this chapter, the goal is only to briefly review the major accounts and summarize their known problems.

§1 The Deductive Nomological Account

In the latter days of logical positivism⁴, explanation appeared to be a spooky and ill-defined notion. Whereas the criteria for being a prediction were clear, and predictions' quality precisely evaluable in terms of experimental accuracy, the criteria for being an explanation were unknown, and no metric for the quality of explanation was apparent. Some linguistic acts seemed to cause people to understand something they had not understood before, but such understanding appeared to be an entirely internal affair, immune to the rigors of scientific testing and observation.⁵ However, it certainly seemed that, in addition to predicting future events, respectable scientists spent much of their time explaining past or present occurrences.

Given this context, the nature of the first formal account of the nature of explanation—Hempel and Oppenheim's Deductive-Nomological (D-N) model—seems almost inevitable. Given the unquestionable scientific adequacy of prediction, Hempel and Oppenheim made the natural move of suggesting that explanations are nothing but post-hoc predictions. At first putting aside fundamentally indeterministic processes, the

⁴This was famously a family of views that (in addition to a particular conception of mathematics and logic) prized observation, verification, and the scientific method. "The synthetic statements of the empirical sciences meanwhile were held to be cognitively meaningful if and only if they were empirically testable in some sense." (Uebel 2012)

⁵"Such expressions as 'realm of understanding' and 'comprehensible' do not belong to the vocabulary of logic, for they refer to the psychological or pragmatic aspects of explanation." (Hempel 1965, 413).
suggestion was that an event was explained when it (or, more technically, a sentence expressing its occurrence) was derivable, and hence definitively predictable, from the right sort of prior information. Moreover, it was accepted by both the (majority of) philosophers and the (majority of) scientists of the time that the universe is governed by certain physical laws. Thus, Hempel & Oppenheim set out to fashion an account wherein statements of universal laws glued together deductions from initial conditions to the occurrence of the explanandum. While it is difficult to pin down exactly what constitutes a law, the general idea is that laws are generalizations that are universally true, exceptionless, contain only general predicates, and support counterfactuals. Newton's laws of motion were paradigmatic examples (aside from their regrettable falsity). An explanation, according to this D-N account, is a logically sound argument whose conclusion is the explanandum and whose premises include a statement of one or more laws of nature, zero or more initial conditions (statements of what the world was like at or before the time of the explanandum), and zero or more boundary conditions (statements that delineate which subsystems of the world the explanation will take into account).

\[\text{(6)}\]

Obviously this characterization is woefully inadequate and underspecified, but it is not at all obvious that there exists a characterization that is satisfactory, particularly given the added methodological constraints imposed by logical positivism. Thankfully this tends not to matter for the purpose of explaining the motivation behind the view or presenting counterexamples; counterexamples typically grant the proponent of D-N whatever laws she needs to get the positive examples of explanation correct, and then demonstrate that these same laws either overgeneralize or were not necessarily exemplified in the actual explanation in the way demanded by the account. Neil Tennant (2008) argues that at least one of the putative counterexamples—that one can explain the length of a pendulum by its period—actually turns at least in part on a hidden misconception regarding what laws were involved in the explanation favored by the D-N proponent (i.e. the explanation of the period of the pendulum in terms of its length). There are however sufficiently many counterexamples that that particular one can simply be excluded, pending a more general demonstration that all counterexamples (or even all counterexamples of the same general type) commit the same error.
Even putting aside general concerns regarding what counts as a law, D-N proved susceptible to both positive and negative counterexamples. Since much of the motivation for my account (see next chapter) is that there is a particular pattern to the counterexamples facing all other accounts, it is worth looking at the counterexamples to the most influential account in some detail. In his overview of the philosophy of explanation, Wesley Salmon (2006, p. 46-50) gathers together a bevy of counterexamples to the D-N account (what follows is his list of examples, with added description. One stylistic alteration is to foreground the role of context, which for reasons of historical accuracy Salmon does not emphasize until later in his volume):

CE-1: "The eclipse" (Salmon 46). The D-N model has no requirement that explanans be temoporally prior to the explanandum, whereas we (at least frequently) intuitively reject putative explanations solely based on the absence of that temporal ordering. (In typical contexts) we accept that a derivation of the moon's position between the sun and the Earth based on the past positions and momenta of those objects counts as an explanation of the eclipse, but (in those same typical contexts) we reject a putative explanation of the eclipse that derives its occurrence from future positions and momenta of the bodies involved. But the deductive structures of both putative explanations are identical, and the exact same laws pertaining to motion of massive bodies and the propagation of light are utilized in both cases. So Hempel does not have the resources to claim that the former explanation is genuinely explanatory whereas the latter is not. (See Chapter 7 for a more extensive discussion of this example and CE-2)
CE-2: "Bromberger's flagpole" (Salmon 47). The D-N model has no requirement that causally prior objects and events also be explanatorily prior. From the height of a flagpole at time t, the elevation of the sun at t, and laws about the propagation of light, one can deduce the length of the flagpole's shadow at t. However, if one were instead given the length of the shadow, the elevation of the sun at t, and the same laws one could just as well deduce the height of the flagpole7. As in CE-1, D-N does not seem to have the resources to differentiate between the intuitively good former explanation from the intuitively malformed latter explanation. Opponents (though, oddly enough, not really Bromberger, who first noticed this counterexample (1966)) typically identify the asymmetry between the flagpole-to-shadow explanation and the shadow-to-flagpole non-explanation as at root a causal asymmetry.

As presented here and elsewhere, the example typically assumes that the flagpole and its shadow exist at exactly the same time, but to a certain extent that is a simplifying assumption. The shadow at any time is really related to the flagpole at a very slightly earlier time (the movement of the light around the shadow is not instantaneous). This particular example then might seem to be a special case of the temporal asymmetry

7It sometimes goes unnoticed in these debates that one could not really do either derivation particularly well at all. To actually deduce, say, the length of the shadow, one would also need to know facts about the curvature of the Earth and probably other light sources nearby. It is at least conceivable that an alternate source could precisely fill in the gap for the light obstructed by the flagpole—however unlikely that scenario, in order to do a proper deduction on must find some way to exclude it. No doubt these matters go untouched because it is assumed that whatever difficulty besets the process of derivation infects both the flagpole-to-shadow direction and the shadow-to-flagpole direction of explanation equally, and so this problem could not possibly help proponents of the D-N model from breaking the alleged symmetry.
brought out by CE-1—any measure that would rule out explanations of earlier events by later ones would also rule out the degenerate shadow-to-flagpole pseudo-explanation.

Whether one can construct a counterexample of the flagpole type that does not turn out to be an example of the eclipse type as well depends on whether one believes there are cases of synchronic- or reverse-causation. If however causes always precede effects, then any constraint on explanation that rules out the possibility of later events explaining earlier ones will, *en passant*, rule out the possibility of effects explaining causes, without needing a separate explicit restriction.

CE-3 "The barometer" (Salmon 47) The third counterexample has the same form as the first two, alleging that D-N lacks the theoretical resources to distinguish between cases that intuitively we judge to have different explanatory merit. From the knowledge that a low-pressure front is coming and a lawful relationship between pressure fronts and rainfall we could deduce that it will rain, and so explain the rainfall. However, we can also deduce that it will rain from the drop in a barometer—both, we can suppose, are caused by drops in pressure in a law-like manner. While structurally slightly different

---

8 Though this assumption is patently somewhat absurd, it is not obviously in any worse shape than the assumption in the good explanation that the low-pressure front causes the rain in a law-like manner. Unlike in the previous examples, the considerations in favor of the paired good and bad putative explanations are not totally symmetrical—in order to be able to derive the rainfall from the barometer by appeals to law one must at some point still invoke the law relating air pressure to rainfall. This point is often missed in the literature, where it is sometimes duly accepted that the direct relationship between the barometer and the rain is lawful. (Salmon, for example, grants that this is the case "for the sake of argument", which is an odd rhetorical move since the proponent of D-N would have no reason to grant that that relationship was lawful. Regardless of the universality of the generalization, it clearly does not support counterfactuals—if it did, and barometer-moving counterfactually supported rain-coming, the case would be radically different from the one intuitively pictured. (If rain-coming counterfactually supported barometer falling, the case would again be different, and at worst analogous to CE-2.) Thus the barometer-to-rain explanation requires the same idealization the pressure-to-rain explanation does,
from CE-2, the upshot is essentially the same—D-N gives the wrong results because it does not have the resources to distinguish the explanatory relevant cause-effect relationship from other, explanatorily inert statistical relationships among phenomena in the same causal web.

CE-4 "The moon and the tides" (Salmon 47). Long before Newton came along and explained how the moon caused the tides, sailors knew that the position and phase of the moon was relevant to the character of the tides. Salmon claims that "given the strict law correlating the position and phase of the moon with the ebb and flow of the tides, it was obviously within their power to construct D-N explanations of the behavior of the tides." (ibid.) However, we would still not want to say that the sailors could explain the tides.

There is another dangerous problem lurking here, which is that Hempel did not make it at all clear what relationship the bits of language actually used in explaining acts must bear to the true laws of nature in order for the act to count as one of explaining. Given that what the sailors articulated likely was not the precisely true law, what relationship would it have to bear to the actual law in order to count as close enough? One might think that the means of escape suggested to Hempel for this sort of counterexample (i.e. that we, but not the sailors, cite the real laws governing the system) merely lets him exploit the vagueness of his account, which is hardly a virtue of the system.

and then yet another idealization. Still, while this is technically an asymmetry, it likely cold comfort to the proponent of D-N: "Your putative but bad explanation requires acceptance of two actually false laws whereas my good explanation only requires one" is not a very satisfying defense.
CE-5 "Syphilis and paresis" (Salmon 49). Suppose cases of the disease paresis are almost always caused by untreated latent syphilis. Nevertheless, even latent, untreated syphilis rarely results in paresis—say it does about 25% of the time. It seems that the mayor's untreated latent syphilis is the explanation of her paresis, yet the D-N account cannot get this result. It demands that explananda be deductively proven; a weakened version called the Inductive Statistical (I-S) account merely demands that the explananda be rendered probable, but even that goal is out of reach in explaining the paresis.

The proponent of D-N could contend that the *ideal* explanation must really make the explananda deducible, and that the invocation of probabilities merely reflects our ignorance regarding what the relevant factors are. There would be two problems with such a reply. The first problem with claiming that the probabilistic paresis explanation is merely a proxy for an ideal deductive explanation is that this response, if fair, could be used to defend any account against any putative counterexample. One could employ this strategy to defend the claim that explanations are really a variety of lawn chair, and that what bits of language we take to be explanatory are only really explaining in virtue of their being a proxy for such chairs. A defense against an apparent counterexample that hinges on the *differences* between explaining acts and ideal explanations is a cheat until we are given an account of what *similarities* there must be. Unless we know what the relation the stated explanation of paresis must bear to the ideal explanation, what reason is there to suppose that citing the syphilis counts at all?
The second problem with claiming that all seemingly probabilistic explanations are really proxies for deductive explanations is that quantum mechanics is an incredibly successful predictive science; given Hempel's motivations, he would thus have wanted to classify it as an explanatory science. But in the case of quantum mechanics the majority opinion is that there are simply no hidden variables to be had—the processes are fundamentally probabilistic. It is easy to construct an example structurally identical to the syphilis and paresis case that one could only claim was really deductive at the cost of going against the considered view of modern science. Even if one happens to think the science is wrong in this case, it seems odd to say that a scientific theory must be either wrong or not-explanatory because it conflicts with a philosophical account of explanation.

CE-6 & CE-7 "Explanatory Irrelevance" (Salmon 51). Extraneous premises do not generally detract from a deduction—if a deduction is sound, then no additional premises will ever make it invalid, and hence no additional true premises can ever make it unsound. Obviously, if a natural law is a premise in an argument, adding more premises will not change its status as lawful. Therefore, given the D-N criteria regarding what counts as an explanation, the addition of true premises to an otherwise good explanation cannot thereby render it a bad explanation. The premises of a D-N explanation are supposed to be what explain the conclusion, yet it seems that once some of the premises do so, we can throw in whatever irrelevant information we want and have it "explain" the explanandum. Salmon points out that on Hempel's account the fact that John is a man and
the fact that he takes birth control are both parts of an explanation of why he did not get pregnant, even though the latter is wholly irrelevant given the former. Worse still, the fact that this grain of salt was hexed can be part of the explanation of why it dissolves in water (assuming that a statement of the salt's hexedness can be put in sufficiently naturalistic language to be true). Yet not only is the hexedness irrelevant to the dissolving, but citing it seems to positively mislead the recipient of such an explanation.

CE-8—The Caveman's Ink Stain. The simplest, but, perhaps, most devastating counterexamples to D-N concern not its sufficiency but its necessity. The fact is that the vast majority of explanations given and accepted in daily life, and even in the sciences, simply do not have a D-N structure. As Michael Scriven (1962) argues, I can explain the ink stain on the floor by citing the fact that I knocked over the ink pot on the desk. I do not generally cite Newton's (much less Einstein's) laws of gravitation, and, more importantly, doing so would in most contexts seem utterly beside the point. We could explain perfectly well to a (English-enabled) caveman what had happened without needing to first explain the behavior of bodies in orbit or other gravitational phenomena. The standard reply to this is that what we say to the caveman is not the real explanation, but merely some sort of epistemically available proxy. The problem with this strategy is that it is not at all clear what relation the explanation actually offered must bear to what Woodward calls the "hidden structure", and he systematically dissects all the various possibilities in chapters 4.3-4.4 of his Making Things Happen. At the risk of trivializing

---

9This does not appear on Salmon's list.
Woodward's complex argument, the result can be put fairly succinctly—if the "hidden structure" of an explanation can be wholly unknown to actual explainers, in a vocabulary wholly distinct from what actual explainers are familiar with, and potentially even computationally intractable for any being that might ever explain, then what work could it possibly be doing?

An examination of these counterexamples naturally leads to the conclusion that D-N lacks the conceptual resources necessary to differentiate between genuine explanations and formally analogous defective explanations; some heretofore banished resource must then be readmitted. However, there are at least two obvious candidates for conceptual resources one might need to employ to distinguish real explanations from formally analogous interlopers.

While this is not clear of all the counter-examples, many of them are defective because they fail to track genuine causal relationships. Thus, the first conclusion one could draw is that the problem with D-N is that it does not employ a notion of causation. Since causation was still considered scientifically suspect (by philosophers), the eventually dominant thread in the literature on scientific explanation came to see itself as having two primary tasks—first to give an account of explanations in terms of causal relations, and second to give an independent account of causal relations.

The second conclusion one could draw from the counterexamples is based not on their content but on the very criteria used to evaluate them. We put forward two examples, and state that somehow one appears or seems or feels explanatory, whereas the
other does not. If we could introduce directly into the theory of explanation an account of this effect that only genuine explanations are supposed to have—which looks conspicuously like the banned notion of understanding—we could rule out the counterexamples very easily. One could then take the task of producing an account of explanation as a two pronged challenge, just like the one facing those who sought a causal account—one must first give an account of explanation in terms of understanding, and then give an independent account of understanding. For some reason this latter route has been explored significantly less deeply than the former.

§2 A Statistical Detour—The Quantum Mechanical Threat

§2.1 Salvaging D-N from Electrons

Even though, looking at all the counter-examples together, the most pressing conceptual shortcomings in philosophical accounts of scientific explanation appear to be the absence of any reference to causation or understanding, it was the failure of D-N to account for statistical explanations (as exhibited in CE-5) that had the most immediate impact. Given Hempel's naturalistic aspirations and emphasis on predictive ability as key to understanding, it is no surprise that the failure of his theory to classify what some scientists took to be perfectly good explanations in quantum mechanics as genuinely explanatory would be problematic. The trouble with quantum mechanics is that it occupies a paradoxical position vis-a-vis the D-N account—it issues in extremely
accurate predictions, but only for sufficiently large sets of measurements. It is typically of no use for predicting specific phenomena or individual measurement results. D-N valorizes arguments' predictive potential and their deductive form—the serious possibility of the former with no hope of the latter revealed a schism in the very meaning of what it is for an explanation to be good.

Faced with the problem of statistical predictions, there are several ways one could attempt to modify the D-N account. First, one could maintain that either quantum mechanics must really be deterministic (and hence at least possibly framed using deductive arguments) or that it is not really explanatory. As already discussed with respect to CE-5, this puts the philosopher of scientific explanation in the awkward position of dictating to the scientist what she should take as explanatory or, worse yet, dictating to the scientist how the world must be. One could of course push such a revisionary line, but since the original motivation for forming an account of explanation was to model and precisify the actual practice of science, such a sharp departure would have been anathema to someone like Hempel. Moreover, there is no apparent principled basis for such a revision—if D-N cannot model successful scientific practice, there is no alternative validation at hand for accepting D-N in the first place.

As a second approach, one could contend that what quantum mechanics really explains are certain probabilistic statements of the form "there is a 50% chance that the electron will be observed with this apparatus". This approach was considered by Hempel under the heading of Deductive Statistical (D-S) Explanation. It is easy to overlook the radical difference between D-N and D-S explanation. The important difference is not that
one is now using probabilistic instead of deterministic laws to explain events, but that one who offers a D-S explanation is no longer in the business of explaining events at all. We cannot explain why this electron was observed where it was. Nor can we even explain why this group of electrons was observed where it was. Most problematically, for any finite number of actual observations we cannot even really explain the fact that we observed the electron with the apparatus 50% of the time—while this outcome was vastly more probable than any other, some other distribution was possible, and so we could not have deductively predicted it.

A third approach to probabilistic explanation, also due to Hempel, is to relax the requirement that the explanans make the explanandum certain to the requirement that they form an argument to the effect that the explanandum was highly probable. Several problems beset this approach. Of most concern to Hempel was difficulty in determining precisely how much information should be contained in the explanans. Since inductive logic is non-monotonic, an argument to the occurrence of some particular event can turn from strong to weak as more information is included in the premises. It would be extremely odd if some occurrences could only be explained when clearly pertinent information is ignored. For example, suppose we have Alex, a rod of $^{266}$Seaborgium (half-life: .36s) that undergoes no measurable radioactive decay over the course of 1 minute—an absurdly unlikely event. If we accept that explaining statistical events means showing that they were probable, Alex's failure to decay should be wholly unexplainable. However, it is relatively easy to produce a cogent inductive argument that has as its conclusion that Alex does not decay over the course of a minute.
1. Alex is a rod.

2. [Here we put a complicated derivation from the laws of nature that most of the rods likely to exist in the universe are highly unlikely to decay in a minute]

C) Alex does not decay over the course of a minute.

The double lines here indicate that the conclusion can be reached with high inductive certainty. This explanation is clearly defective, however—the conclusion has only been given the veneer of likelihood by ignoring the central bit of information that Alex is composed of $^{266}$Seaborgium. In order to prevent such "explanations" from counting, there needs to be some requirement that one not gratuitously exclude information, but this requirement proves surprisingly difficult to formulate precisely. It will not do to require that all available information be accounted for in the explanans, because in most circumstances an explanation is offered only once we know that a particular event occurred—including that bit of information would of course make the argument trivial (and deductive). Hempel instead invoked the Requirement of Maximum Specificity, which (roughly) requires that one include all information in one's explanans that one knows or could deduce affects the probability of the explanandum occurring, except that information which logically entails it (Hempel 1965, 399-400). This solves Hempel's concern, but only at the cost of making all statistical explanations necessarily constrained to a particular epistemic perspective.
The larger problem with the I-S account though is the same as the problem with the D-N account—when compared to our pretheoretic intuitions regarding what is and what is not explanatory, it simply gets the data wrong. To the extent that we can explain probabilistic happenings at all, we seem just as able to explain events whose occurrence was unlikely as events whose occurrence was likely. Given two radioactive atoms and arguments to the effect that one has a .01% chance to decay and the other has a 99.99% chance to decay within a given time period, if both actually do decay the former seems as well explained as the latter. Both had a certain probability of decaying, and both did—if the processes were really indeterministic, there is no more mystery in one case than the other. Given Hempel's allegiance to the principle that explanation and prediction are symmetrical he would deny that the two explanations are on a par, but outside of these theoretical commitments there is little reason to do so. (Again, I contend that what drives intuitions here seems to be that both explanations provide all the understanding that can be provided.)

§2.2 Statistical-Relevance Explanation: Drifting Towards Causation

At first, philosophers such as Wesley Salmon, Richard Jeffrey, and James G. Greeno (1971) diagnosed the main shortcoming of the I-S model of explanation as its failure to really grapple with the fact (or at least strong possibility) that some events in the world are governed by genuinely stochastic laws, and that sometimes, in such
situations, improbable results come to pass. Their proposed remedy was to claim that explanation of such events was not in the business of providing arguments at all, but rather of merely specifying the factors that were statistically relevant to the occurrence of that event. They thus proposed the Statistical-Relevance (S-R) model of explanation. An ideal S-R explanation would be a partition of some larger contrast class into sub-classes ("cells"), in virtue of all and only those differences that affect the probability of the explanandum occurring. The idea is that, by so doing, one puts forward all and only those factors that are relevant to producing the event in question. When we know that Albert was a member of a certain gender, a certain age group, and had a particular type and location of upbringing, that people in that cell have a certain probability of stealing cars, and that within that cell there are no other statistically relevant factors to account for who steals a car and who does not\textsuperscript{10}, then we have done all we can to explain why Albert stole the car.

The main problem with this account is right in its name—the sort of relevance it highlights is statistical relevance, and that alone. The problem with this is that, while statistical and causal relevance often overlap, where they diverge we generally only consider causal relevance explanatory. This problem crops up early in Salmon's exposition of an example meant to demonstrate his view:

\textsuperscript{10}Of course, there will always be at least one further partition—if we further partition Albert's group (G) into those who steal cars (G\textsubscript{1}) and those who do not (G\textsubscript{2}), the probability that Albert stole the car will of course be markedly different depending on whether he was a member of G\textsubscript{1} or a member of G\textsubscript{2}. Getting around this sort of concern (and more subtle variants) requires extreme care in stating the conditions on what is an acceptable partition for explanatory purposes, as well as what it means to say that an event that definitively did (or did not) occur had some probability other than 1 (or 0). Since these caveats have no bearing on the criticisms to be presented, I simply grant for the sake of argument that such details can be worked out satisfactorily.
Hempel's example of quick recovery (Q) from a strep infection provides a different illustration. In this example, the original reference class is the class of people who have streptococcus infections (S). Clearly, the partition of this class into those who receive treatment with penicillin (T) and those who do not (~T) is relevant. Also, a partition can be made into those who have penicillin-resistant infections (R) and those whose infections are not penicillin-resistant (~R). If, however, we combine these two partitions to form the cells the resulting partition is not a relevant partition, for the probabilities are not all different. If the infection is penicillin resistant, it makes no difference (I assume) whether penicillin is administered or not, and (I further assume) the probability of quick recovery is the same for a person with a non-penicillin resistant infection who does not receive penicillin as it is for anyone with a penicillin-resistant strain. —Salmon (1989, 64-65, emphasis mine.)

Reading this, I found the first assumption rather at odds with my own intuitions. Even if someone with regular strep who goes untreated, someone with penicillin-resistant strep who is given penicillin, and someone with penicillin-resistant strep who is not given penicillin all have the exact same probability of recovering, I would certainly not want these various possibilities lumped together in an explanation of my non-recovery. Consider the dialogue that would result if a doctor took seriously Salmon's view: "The reason you have not recovered is either that you had penicillin-resistant strep, or that you
had perfectly treatable strep and I did not give you penicillin—as you know, it is 
explanatorily irrelevant which." The partition invoking only those factors that make a 
*statistical* difference is thus too coarse-grained to capture all the factors that intuitively 
make an *explanatory* difference. Hugh Lehman, in his 1972 "Statistical Explanation",
argued that there are also cases where it makes too fine grained a distinction. Suppose 
Jones were to die of a ruptured spinal cord but, as it happens, he was having a heart attack 
at the same time. The odds of Jones dying given that he has a ruptured spinal cord and a 
heart attack is, we can suppose, higher than the probability of his dying from a ruptured 
spinal cord alone. Thus, an S-R explanation would include a partition of individuals with 
ruptured spinal cords into those with heart attacks and those without—but, given that we 
know that the heart attack was causally irrelevant to Jones's death, we also tend to think 
that it was explanatorily irrelevant.

The impossibility of merely statistical relations obviating the need to appeal to 
causes in daily-life was forcefully made by Nancy Cartwright in her 1979 "Causal Laws 
and Effective Strategies". She argues that which statistics we take as relevant depends on 
preexisting beliefs about possible causes and causal mechanisms, and that these latter 
beliefs are essential for planning our lives. While Cartwright does not focus on 
explanation, it is clear that her conclusions apply to merely statistical explanations. She 
cites the example (a variant of Simpson’s paradox) of a study that demonstrated that 
women were getting into graduate school at a lower rate than many. However, the 
apparent discrimination dissipates if one examines the rate at which women are admitted 
to any particular department—as it turns out, women tend to apply to more selective
programs. The fact that partitioning women by department reverses the troubling trend seems to explain the original fact. Suppose, however, that, as it happens, women who roller blade have a higher rate of acceptance to graduate programs than men who roller blade, and women who do not roller blade have a higher rate of acceptance than men who do not roller blade. If proportionately more women roller blade and graduate schools happen to be\(^{11}\) more selective with respect to roller bladers than non-roller bladers, it could still be the case that women get into graduate school at a lower rate than men. But the partition of women into roller bladers and non-roller bladers seems to do nothing to explain the originally perplexing datum, despite this explanation's being statistically equivalent to the legitimate explanation in terms of partition by department applied to (Tables 1 and 2). The difference, Cartwright argues, is that only in the former case does the statistical relation signify anything causally relevant. If the latter sort of partition were accepted as an explanation, we could always find some convenient way to explain any fact that might otherwise be taken as evidence of a deep underlying problem.

\(^{11}\) Note the causal structure of the world implicit in the phrase “happened to be”.  

26
Table 1 Not Worrisome Partition of Graduate School Applications

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Psychology—</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Application Rate</td>
<td>25%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>(500,000)</td>
<td>(1,000,000)</td>
</tr>
<tr>
<td><strong>Psychology—</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptance Rate</td>
<td>55%</td>
<td>50% (500,000)</td>
</tr>
<tr>
<td></td>
<td>(275,000)</td>
<td></td>
</tr>
<tr>
<td><strong>History—</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Application Rate</td>
<td>75%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>(1,500,000)</td>
<td>(1,000,000)</td>
</tr>
<tr>
<td><strong>History—</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptance Rate</td>
<td>15%</td>
<td>10% (100,000)</td>
</tr>
<tr>
<td></td>
<td>(225,000)</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>500,000</td>
<td>600,000 (30%)</td>
</tr>
<tr>
<td></td>
<td>(25%)</td>
<td></td>
</tr>
</tbody>
</table>
Table 2 Still Worrisome Partition of Graduate School Applications

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Non-Roller Blader</strong>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Application Rate</strong></td>
<td>25%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>(500,000)</td>
<td>(1,000,000)</td>
</tr>
<tr>
<td><strong>Non-Roller Blader</strong>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Acceptance Rate</strong></td>
<td>55%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>(275,000)</td>
<td>(500,000)</td>
</tr>
<tr>
<td><strong>Roller Blader</strong>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Application Rate</strong></td>
<td>75%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>(1,500,000)</td>
<td>(1,000,000)</td>
</tr>
<tr>
<td><strong>Roller Blader</strong>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Acceptance Rate</strong></td>
<td>15%</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>(225,000)</td>
<td>(100,000)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>500,000</td>
<td>600,000</td>
</tr>
<tr>
<td></td>
<td>(25%)</td>
<td>(30%)</td>
</tr>
</tbody>
</table>

Tables 1 & 2: Two statistically isomorphic partitions of 2,000,000 women and 2,000,000 men's graduate school application and acceptance rates. While the partition in 1 seems to explain the disparity found in the final row, the partition in 2 does not seem explanatory at all.
As with the counterexamples to D-N, the counterexamples to S-R hinge on the theory's inability to differentiate between putative explanations that we take to be explanatory and those we do not. The conclusion is the same—S-R simply does not have the conceptual resources necessary to distinguish explanations from non-explanations. In the case of S-R, the difference between explanatory partitions and non-explanatory partitions seems to be that the former, but not the latter, only take into account genuinely causal statistical relations.\(^\text{12}\)

It is worth, flagging, once again, that all the things that seem to be genuine explanations differ in one other way from those things that do not seem to be explanations—namely, they differ in that they seem different. Aside from causation, one route that could have been explored would have been to incorporate the very difference in effect by which we judge something to be explanatory or not into an account of explanations. That is what I do in the next chapter.

§3 Causation Proper

Problems with the S-R account of explanation forced philosophers to take seriously the possibility that real explanations of a phenomenon are those that give

\(^{12}\)The quantum cases are often tricky here—if one alters an apparatus so that the probability of a particle being observed in a certain position changes, is this really a causal difference? The answer depends on whether one thinks there can be probabilistic causes, which question is well beyond the scope of this essay. Putting the word 'cause' to the side, however, it does seem difficult to dispute that what one has done has somehow affected the outcome. Cartwright's approach is perhaps useful here—if one wants to observe a particle in a particular location, performing various manipulations are an effective strategy for doing so, in a way that giving up one's roller blades are not an effective strategy for getting into graduate school.
information about that phenomenon's causes. For most people pursuing this route, the project then became one of producing a naturalistically respectable account of causation. Such accounts are open to objection on two fronts. First, one could examine the offered account of causation and argue that it fails, be it because it does not capture causal judgments, is not naturalistically respectable, or just because it is somehow trivial or uninteresting. Alternately, one could grant that a given philosopher has offered a good account of causation, but argue that this is still not all there is to explanation. I pursue this latter strategy.

Jaegwon Kim, in his 1992 "Explanatory Knowledge and Metaphysical Dependence", points out that one event can explain another, without the connection between events being obviously causal. For example, we can explain the event of Xanthippe's becoming a widow by pointing to the event of Socrates's death, but did the latter cause the former? (Kim 1992, p. 67) On some accounts of causation, it could not have done so. For example, on Wesley Salmon's naturalistic accounts, according to which effects can propagate from an event no faster than information, the link between the death and the widowhood cannot be causal. (It would be bizarre if, after Socrates died, Xanthippe was still married for a few microseconds until the light from his death happened to reach her environment. It seems more reasonable to say that her widowhood was simultaneous with his death, at least provided that they are moving in close enough to the same inertial reference frame for talk of simultaneity to be meaningful.\textsuperscript{13}) Kim

\textsuperscript{13}Factoring in general relativity, it is also required that they be close enough for the curvature of space to be negligible.
suggests that, instead of causation, explanations invoke metaphysical dependence relationships. One problem with this proposal is that if there is any notion in philosophy more vexing to make precise than *causation*, it is *metaphysical dependence*. Kim does not attempt to do so. In perhaps the best known paper where *metaphysical dependence* is tackled explicitly Gideon Rosen, interestingly, treats such dependence as a necessarily synchronic relation. However, rather than incorporating causation into a broader category, treating explanation as a matter of metaphysical dependencies so defined actually rules most causes—which are (at least) typically diachronic—out entirely! This is not a devastating objection—one could formulate a different account of metaphysical dependence or, less elegantly, simply say that explanations point to either causes or metaphysical dependency bases—but it points to some of the unseen tangles that can envelop one who attempts to traverse this road.

Even within the physical sciences, there are things that seem to be perfectly legitimate explanations that nevertheless do not appear to point to causes. Eliot Sober (1983) discusses "equilibrium explanations", where the end macro state of a large population is explained by the statistical facts that render such a state uniquely likely or stable. Suppose, for example, we are playing a coin-flipping game that I win if the coin comes up heads, but, for reasons unknowable, I get ten chances. When the coin shows heads on the third flip, a passerby asks why I won—the proper explanation seems to be that, given the number of chances I got, it was overwhelmingly likely that heads would come up at least once. By contrast, it seems wholly beside the point to go into details regarding the causal processes that led to the third flip falling out as it did.
To take another example of non-causal explanations in the physical sciences, biologists will often explain a phenomenon in terms of the mechanism that produces it. No doubt this explanation involves causation, but it also involves elements that are not obviously causal—for example, hierarchical decomposition into parts. Explaining a mechanism as simple as a cell's ability to regulate ions requires accounting not just for any one particular causal process, but for how a range of processes working in parallel are situated precisely so as to achieve the result in question.

Evaluating the role of causation in explanations in human sciences is tricky, as the success of a causal model of explanation will depend on how much of human action can be reduced to causal terms. If intentions, goals, beliefs, desires, or plans cannot be readily made sense of in purely causal language, then the possibility of good explanations being framed in such terms should worry the theorist who claims that all explanations must be causal. It is thus likely that causation is insufficient to account for all explanation in fields like psychology and sociology.

The biggest limitation to causal accounts is their inapplicability to explanation in fields interested in the examination of necessary truths. Mathematics most prominently, but also philosophy and perhaps other fields as well, frequently seem in the business of explaining one necessary truth with another. For example, it is indisputable that mathematicians sometimes take some things—typically a subclass of proofs—to be explanatory (see e.g. Hafner & Mancosu 2009). However, all accounts of causation generally agree that, if A causes B, then B would not have happened if A had not had happened. Thus, it makes little sense in mathematics to talk about causation—what
would it mean for a proof not to have happened? The proof never did happen, and, even if it had, it would have had to have happened. The proof does not cause the result it seeks to explain; thus, if mathematicians are ever right to consider some proofs explanatory, then causation cannot be all there is to explanation. James Woodward, one of the foremost proponents of a causal theory of explanation, acknowledges as much in his *Making Things Happen* (2003); perhaps in response, he restricts his domain of interest to causal explanation (Chapter 1). He then conjectures that a leading account of mathematical explanation—that of Marc Steiner—might be analogous to his own (Woodward, Chapter 5.9). However, this analogy is not explored in depth, and what similarities there are seem superficial.14

Explanations in philosophy are similarly problematic for the proponent of a causal theory of explanation. If utilitarianism is the correct account of morality, then the fact that Johnny's action is not conducive to general happiness seems to explain why his action is wrong.15 More generally, philosophers are regularly in the business of explaining some phenomenon or other—why words mean what they do, why certain logical theorems hold, what the relation is between mind and body, what knowledge is, etc. Sometimes, these explanations might be causal—it is reasonable to suppose that an explanation of why our words mean what they do will involve a causal history of how they came to be

---

14 Regrettably, a more thorough exploration of explanation in mathematics is beyond the scope of the present essay. For a counterexample to Steiner’s view in particular, see Hafner & Mancosu (2009).

15 To avoid a potential concern, it is worth noting explicitly that the claim that ethical theories are explanatory does not fly in the face of Gilbert Harmon's (1988) famous argument that ethical conjectures are not on a part with scientific conjectures because they are not explanatory. His argument regards a narrower question—whether ethical facts play any role in explaining the observations that lead us to posit their existence. As far as I know, he is silent on the question of whether they can explain anything else.
taken as possessing those meaning. However, an explanation of what, for example, knowledge is is unlikely to hinge on any causal facts ("once upon a time a cave-person had the first justified true belief..."). This leaves causal accounts of explanation in the awkward position of being wholly incapable of explaining what a causal explanations is.

It might be the case that, armed with a sufficiently broad notion of causation, the proponent of a causal account of explanation could find ways to argue that equilibrium, mechanistic, and psychological explanations all really point to causes of the explanandum. However, the prevalence of explanations in mathematics and philosophy is not so easy to accommodate. Assuming putative explanations in those two fields are not causal, there are three possibilities: 1) Mathematicians and philosophers to the contrary, there are no explanations of mathematical facts, and surprisingly few explanations within philosophy; 2) 'Explanation' is an ambiguous term, systematically picking out one thing in mathematics, another thing in the sciences, and sometimes one and sometimes the other in philosophy; or 3) Causal explanation is at best an important subclass of explanation generally. Most philosophers would be loathe to accept the philosophical imperialism required to deny mathematicians (and, perhaps, some other philosophers) the right to find explanations where they may. That leaves the proponent of the claim that explanations simply are causes (or statements of causes) with no recourse but to claim that 'explanation' is ambiguous. While ambiguity of 'explanation' is a live option, there are reasons to eschew that route. The most obvious reason is theoretical elegance. Aside from that, while there is no definitive test of whether a term is ambiguous, none of the heuristics commonly used give any reason to suspect 'explanation'. For example, "There
are explanations in physics and mathematics" seems felicitous in a way "Dracula and the Cleveland Indians both have bats" does not.

The correct conclusion, then, is #3—causation might play an important role in the explanation of many physical phenomena, but appeal to causation is not a necessary component of explanation. Note that this conclusion is not wholly negative: given the prominence of causation in many physical explanations and the success of causal theorists of explanation in many domains of interest, it is a constraint on a good theory of explanation that it say why causation should be so important for explanatory ends.

§4 Unificationist Accounts: Drifting Towards Functional Characterization

Michael Friedman, concerned with the inability of extant theories of explanation to elucidate how explanation produces understanding, constructed a new account in his "Explanation and Scientific Understanding". Friedman claimed that "science increases our understanding of the world by reducing the total number of independent phenomena that we have to accept as ultimate or given." (1974, 15) Explanations function by pointing to more general laws and regularities from which we can deduce all phenomena to be explained. For example, statistical mechanics is explanatory because, once we know its laws, various laws pertaining to the behavior of gasses and heat can be derived rather than stipulated. (Friedman requires that the explaining law be acceptable independently of the phenomena it explains--a necessary stipulation, as otherwise a long conjunction of everything a person happened to know would be maximally explanatory. Cashing out
what this independent acceptability is a difficult project, and it is not clear Friedman wholly succeeds (see Salmon 1989, 97-98), but such concerns can be bracketed for present purposes.) Philip Kitcher (1989) followed Friedman with a more complex view that shared with Friedman's the central notion that what explanations enable us to do is to be able to deduce more facts but require fewer resources to do so. On Kitcher's model, the resource to be minimized is the number of argument patterns required to be able to deduce our whole store of knowledge, where "argument patterns" are (roughly) annotated argument schemata.

Two features of Friedman and Kitcher's views were particularly radical. First, the mere recognition that understanding is not merely a byproduct of, but rather a guiding ideal for, the giving and receiving of explanations marked a sharp departure from most earlier accounts. Second, Friedman and Kitcher proposed that whether something qualifies as an explanation might depend not just on the structure and content of that particular explanation, but rather on its effect on a thinker's global cognitive store. This latter point had far reaching implications. If whether something is an explanation can only be evaluated in relation to particular individuals' minds, then the entire project of giving acontextual criteria for determining whether something is an explanation was fundamentally misguided—the exact same thing could be an explanation for one person but not for another.

In other ways, however, Friedman and Kitcher both stayed well within the Hempelian tradition. First, neither used understanding as a necessary constraint on explanation. Rather, they built accounts of explanation that had a particular effect, and
asserted that understanding just was that effect. The particular effects identified—
unification of knowledge into fewer underived laws or argument patterns—seem
plausible candidates for what understanding is. However, it is not clear what could be
made of their views if, upon independent examination, we conclude that understanding is
not wholly identifiable with such unification. Since understanding is no explicit part of
the account of explanation, it at least seems possible that the particular type of things
picked out by unificationist theories as explanations need not cause understanding. Paul
Humphreys, in the perspicuously titled "Greater Unification Equals Greater
Understanding?" argues that unificationist explanations are not in general conducive to
understanding. He points out that, by the criteria Friedman and Kitcher give, a one-axiom
formalization of propositional calculus should provide greater understanding than a three-
axiom formalization, but that just seems obviously wrong.\(^{16}\)

The second way Friedman stayed within the Hempelian tradition—or, in this case,
reverted to that tradition in a way the proponents of causal theories had rejected—is his
requirement that one be able to deduce the explanandum (or a statement thereof) from the
explanans. James Woodward argues that, rather than presenting a wholly new account,
unification theorists were just offering restrictions on the original D-N schema (2003,
Chapter 8). This might help provide D-N the resources to respond to objections to the
effect that it does not give sufficiency conditions for identifying explanations, but not all

\(^{16}\)The example is worked out in detail with respect to Kitcher’s account of unification. Humphreys says
that the same result applies to Friedman, though he does not really discuss whether the single-axiom
formalization meets the “independent acceptability” criterion (what makes an axiom acceptable?).
However, Kitcher elsewhere (1976, especially p. 209-10) argues that, if that criterion is considered as at
all stringent then it rules out many genuine explanations which require conjunctions of laws from
different theories.
the objections are of this sort. By maintaining the requirement that the explanans be formally related to explanandum, unificationist views rule out the vast majority of utterances offered under the heading of explanation from counting as actual explanations. This conviction of mass error calls out for some explanation of the connection between 'explanations' proposed by real people and theoretical explanations, and none seems to work. (Peter Railton (1980), presenting perhaps the most sophisticated version of such a "hidden structure", proposes that things humans call 'explanations' are really just utterances that give some information about the real explanation. However, even he acknowledges that this view has some unsatisfactory consequences. For example, a report on the number of characters the real explanation would possess when written in English would count as an 'explanation' in the colloquial sense.)

§5 Pragmatic Accounts: Drifting Towards Understanding

Taking on board the insight that explanations are best individuated not by their content, but by their role in actual human endeavors, in the 1980s philosophers developed accounts that jettisoned specific structural requirements for explanations in favor of requirements based on role served. The two most well-known accounts in this tradition belong to Bas Van Fraassen and Peter Achinstein. My own account, put forward in the next chapter, also falls into this general category.

Van Fraassen's account, while a pragmatic account in name, carries from Hempel the idea that explanations are in the business of showing why a specific event was to be
expected. Specifically, a statement A explains event E relative to a contextually supplied contrast class X if the ratio \( P(E/A):P(x_1 \text{ or } x_2 \ldots \text{or } x_n) \) (for all \( x_i \) elements in X other than E) is greater than \( P(E):P(x_1 \text{ or } x_2 \ldots \text{or } x_n) \). That is, an explanation favors the explanandum against some set of contextually determined alternatives. However, it intuitively seems that not all explanations require that the explanans probabilistically favor the explanandum. Suppose there is a particle with a \( \frac{1}{3} \) chance of decaying in a given time period, and you ask for an explanation of why it decayed (as opposed to the obvious contrast class of it not decaying). As it happens, I had just fired a ray at it designed to shore it up, such that there was a 99% chance the ray would have insured that it did not break up at all. However, there was also a 1% chance that the ray would just blow it to pieces immediately, and that's actually what happened in this case. Pointing out that I just fired the ray and, quite improbably, it exploded the particle seems to be the explanation of why it decayed, though the information that I fired the ray actually made the explanandum significantly less likely to have occurred.

Achinstein's account is complicated, and turns largely on a careful analysis of English grammatical forms that might not be applicable to other languages in full generality. Most of what he says is plausibly correct: he argues that the process of explaining is conceptually prior to the product explanations, that one can only explain via intending to create understanding and that what constitutes understanding will depend on

\begin{footnote}
\footnotesize
Of course, the fact that I fired the ray and the particle decayed does probabilistically favor the particle decaying, but it clearly will not do to say that I fired the ray explains the decay because of its relation to the fact that I fired the ray and the particle decayed. The fact that the circus was one town over at noon bears the same relation to the fact that the circus was one town over at noon and the particle decayed, but the circus being one town over at noon clearly explains nothing about particle decay.
\end{footnote}
both immediate and broad historical context (including questions of emphasis and broader scientific context). On his view, one explains by intentionally uttering a sentence with the right content, emphasis, and intentions. The "right content" requires that the explanation provide a "complete content-giving" answer to the question of why the explanandum occurred. To take Achinstein's stock example, we explain Nero's fiddling by providing "The reason Nero fiddled was that he was happy" as the correct answer to the question "why did Nero fiddle?" (1983, p. 72—see Chapter 3, §4.3.2 below for more discussion) However, there is an obvious problem with this view, as pointed out by Salmon:

This view of explanation seems seriously question-begging...what objective relationship must obtain between the fact that Nero was happy and the fact that he fiddled to make "The reason Nero fiddled is that he was happy" a correct answer?

--Salmon (2006, 148)

Salmon's point can be brought out even more sharply when we note that, by the criteria Achinstein gives, "what explains that Nero fiddled is that he was happy" also, when uttered with in the appropriate intention in the right context, count as an explanation of Nero's fiddling. Salmon's point then is that Achinstein's theory of what it takes for one fact to explain another boils down to the fact that it does so, which is not especially illuminating. (For a more careful reconstruction of Achinstein’s account see Chapter 3, §4.3.2)

However, the fairness of Salmon's charge that Achinstein's view is "question-begging" of course depends on precisely what question Achinstein was trying to answer.
If his goal was to elucidate what explanation is, and in virtue of what we can honestly say that something is an explanation, then it seems that we have made very little headway. The title of his book (The Nature of Explanation) and certain passages might lead one to believe that is his goal. However, if instead his account is viewed not as a pragmatic account of explanation—that is, as an account of explanation that grounds the nature of explanation in pragmatic considerations—but rather as an account of the pragmatics of explanation—that is, as an account that illustrates the role explanations play in actual pragmatic contexts—then everything he says is both plausible and illuminating. Since, however, the purpose of this essay is to explore accounts of what explanation is, the latter reading of Achinstein’s account can be fairly bracketed.

§6 Conclusion: Coming Ashore

Deductive-Nomological and causal accounts of explanation, while plausible in certain domains, are not sufficiently fine-grained to differentiate between what we intuitively take to be explanatory and what we do not. I conjecture that this is because our intuitions are informed not just by what the form and content of a particular explanation is, but by the effect it has. Unificationist accounts first began to explore the role of explanation in a broader cognitive framework, but were, I argue, too beholden to the Hempelian doctrine of explanation-as-deductive argument to account for the diverse ways people explain things. The pragmatic accounts, which identify explanations largely by their role, are, then, the most promising route. However, both such accounts discussed are, I argue,
either wrong or unhelpful in their details. Baas Van Fraassen's account carries over from Hempel the view that explanations must in some way make the explanandum expectable, which simply is not true of everything in the set of things we take to be explanations. Achinstein's account, by contrast, moves a bit too far in the direction of allowing explanations free form, structure, and role, to the point where it is not clear whether he provides any guidance regarding what explanations actually are. This suggests a middle pragmatic course, according to which explanations are identified by their role, but that role is cashed out precisely in non-circular terms. This is the structure of the account developed in the next three chapters.
Chapter 3: Functional Explaining: A New Approach to the Philosophy of Explanation

Beginning with Hempel & Oppenheim's (1948) contention that explanations are to be identified with a particular formal structure, the general tendency in the philosophical literature on explanation has been to elucidate explanation in terms of progressively more intricate structures or (later) privileged types of content. Michael Friedman (1974) complicated Hempel's identification of explanations with isolated arguments by instead identifying explanations with arguments within a global structure of arguments; Philip Kitcher (1989) made matters still more complex by arguing that explanations were instead to be identified with argument patterns among schemes of argument patterns. Accounts arose that identified explanations with causes or representations thereof. These began as somewhat intuitive process theories when first formulated by Wesley Salmon (as described in his 1989), but quickly became highly technical as well, beginning with the Bayesian-style account produced by Judea Pearl with respect to economics. (This was brought into the philosophical fold by James

---

18 Hempel (1965) later placed an explicit restriction on content, demanding that D-N explanations invoke genuine laws. As of the 1948 paper, however, Hempel and Oppenheim remained optimistic that laws could be at least largely picked out by structural and syntactic markers.
Woodward (2003)\textsuperscript{19} The most recent models of explanation go so far as to treat explanations as systems of dynamical equations (e.g. Zednik 2011).

While the mere complexity of a theory is itself no mark against it, such complexity suggests that perhaps we are adding epicycles to a fundamentally misguided approach. Explaining is something most people seem to have an intuitive grip on; the goal should be to construct as simple and unified a view as possible, consistent with people's judgments regarding what is and is not explanatory, and also with whatever platitudes are regarded by competent speakers as capturing the important characteristics of explaining. One such platitude is that explanations tend to cause understanding. Within the literature, this platitude is frequently either ignored or treated as a corollary of a broader view. My suggestion is that we reverse this sequence, and regard this particular platitude as presenting that which is constitutive of explanation. In §4, I argue that, relative to extant accounts, not only is an understanding-based account simpler, but its extension actually does a better job of accommodating people's judgments regarding what is and is not explanatory.

That there is some close connection between explanation and understanding has never been in serious doubt. Historically, however, understanding has rarely been employed for the purpose of analyzing explanation. Where explanation seemed ever on the verge of being fully reduced to some more tractable notion, understanding has from the first been treated as either a merely phenomenological state of limited scientific

\textsuperscript{19} Woodward's presentation of the view was, admittedly, extremely approachable and non-technical. Issues pertaining to his view are explored in §5.2.
interest or a mere byproduct of explanations (this history will be discussed in more detail in §2). I contend that this gets the cart before the horse. Understanding is not scientifically uninteresting; nor is ‘understanding’ a mere name to be applied to whatever explanations might happen to produce—the quest for understanding is often what the production of explanations is for.

My central thesis is this: explanations just ARE those sorts of things that, under the right circumstances, bring about understanding. This raises the question of why such a seemingly simple account of explanation, if correct, would not have been identified and agreed upon decades ago. The answer is that only recently has it been made possible to analyze explanation in terms of understanding without the risk of collapsing both to merely phenomenological states. For the most part, theories of explanation were for fifty years held hostage to the historical accident that they far outstripped in sophistication corresponding accounts of understanding. A subsidiary thesis of this paper is that advances in the philosophy of understanding have recently made it possible to articulate an illuminating account of explanation in terms of understanding. The ambition of this chapter is thus limited and somewhat deflationary: to demonstrate that, once we can invoke understanding, it becomes plausible that the nature of explanation is rather thin—the interesting philosophy is then in investigating the nature of understanding.

The connection between explanation and understanding will be articulated more precisely in the next section. In §2 I examine why such an intuitively plausible account of explanation was eliminated from consideration, and then in §3 argue that this elimination was a mistake. Looking at counterexamples to historically important views, in §4 I show
that our judgments regarding what cases would or would not cause understanding enable us reliably to distinguish genuine from pseudo-explanations better than other accounts of explanation. At this point, I take the main goal of the paper to be completed— inquiry regarding the nature of explanation will have been redirected from its 60 year course to a new examination of understanding. However, it would be natural then to wonder whether such an examination of understanding is likely to be fruitful. I conclude with a section giving an abbreviated version my own account of understanding, and illustrating how it can shed light on our judgments regarding what is explanatory.

§1 Explanation, Explaining, and Understanding

Let us turn to the actual account of explanation. I begin in schematic form, in order to illustrate the explanatory strategy I pursue, then fill in more details below.

For the most part, most philosophers theorizing about explanation (eg. Michael Friedman (1974), Jaegwon Kim (1992), Paul Thagard, and Hempel & Oppenheim) would have accepted that, *ceteris paribus*, the following traditional claims hold:

TC₁: What makes something a state of understanding is its being that state one is in after being receptive (in the right sort of way) to a good explaining-act.

---

20 Developed in greater detail in chapter 5.
21 Personal correspondence.
22 For a discussion of Hempel & Oppenheim on the link between explanation and understanding see §3.
23 It is not always clear just what the right sort of way is, and thus it is difficult to specify more precisely the exact causal connection these authors envision between explanation and understanding. (This
TC2: What makes something an *explaining-act* is its being a communicative action in which a good/correct *explanation* is conveyed in the right sort of way.

Explanation—standing at the foundation of the traditional edifice—now stands in need of independent elucidation; thus the chain of philosophers of science beginning with Hempel & Oppenheim and continuing to the present day. The contention of this essay is just that, while the links posited in TC are real, its formulation gets the explanatory order precisely backwards:

UBEE1 (Understanding Before Explaining & Explanation): What makes something an *explanation* is its being content which, if true,\(^\text{24}\) could, under appropriate circumstances,\(^\text{25}\) be used in *explaining-acts*.

UBEE2: What makes something an *explaining-act* is its being a communicative action which, when properly internalized by the audience, causes *understanding*.

---

\(^{24}\) The weak requirement that an explanation only need to usable in an explaining-act *if it were true* is inserted to make sense of the following perplexing intuitions: we would say of a theory that we know to be false but which would account for some phenomenon that it is still, in some sense, a (bad) explanation. However, someone who presents that very theory to explain an occurrence of that phenomenon seems not to have explained it at all.

\(^{25}\) See later in this section for a discussion of the force of this modality, though a full accounting of the pragmatics of explanation-talk is beyond the scope of this essay.
The foundation of this edifice is understanding, rather than explanation. I argue in the next two sections that, while this sequence might historically have rendered UBEE a non-starter, it now serves to strengthen it.

UBEE₁—the claim that explanations are just those things that can be offered in explanatory acts—stands in little need of defense, if we assume that UBEE₂ provides a fully satisfactory account of explanatory acts. While proponents of other accounts might well deny UBEE₁, such denial would likely stem in no small part from their concerns regarding the possibility of producing an independent account of explaining-acts. It would be perverse to suggest that, given a satisfactory account of explaining-acts, one nevertheless needs to begin anew when producing an account of explanations. There is really only one fact about explanations that has never been in serious dispute: explanations are the sorts of things we convey, by means of some representational apparatus, from one person to another via communicative acts. The precise nature of the conceptual link between explanations and explaining-acts might not be self-illuminating, but the existence of such a link has never been doubted. Given a satisfactory account of the one, it would be foolhardy not to use it to analyze the other.

Let us say that something is an explanation if and only if it is possible for it to be the content of an explaining-act for some cognizer (the assumption that explanations are composed of representational content is discussed below). Obviously a large degree of vagueness infects such a definition: does a proof with $10^{10,000,000,000}$ steps, which could be used only in an explaining-act among gods, count as an explanation? One suspects the answer to such questions regarding what "counts" as an explanation will vary
contextually, both between speakers and for the same speaker at different times; for present purposes I put these issues to one side.

The real philosophical action, then, is involved in the precisification and defense of UBEE\textsubscript{2}, to which we now turn. We begin with slightly more precision concerning the link between explaining acts and understanding:

FE (Functional Explaining): Explainer B \textit{explains} (engages in an explaining-act with respect to) proposition p to audience A if and only if B produces a representation R that, when properly internalized, causes improved understanding of p in A (where the content of that understanding is similar to R), and B intends to explain\textsuperscript{26} or cause understanding via this production.

FE treats 'explains' as a success term denoting an intentional action. One can then use FE to define non-successful and non-intentional uses of 'explains' derivatively.\textsuperscript{27} I believe the suggested reading of 'explains' best accords with actual linguistic practice.\textsuperscript{28} For present

\begin{itemize}
\item \textsuperscript{26} Since the word "explains" appears within the scope of an intentional operator, its use does not make the definition at all circular.
\item \textsuperscript{27} If one thinks 'explains' is not a success term, then 'explains' just means 'engaged or attempted to engage in an explaining-act as defined by FE'; the remainder of this discussion could then be construed as providing an account of 'successfully explains'. Similarly, if one thinks 'explains' does not necessarily denote intentional action, then the last clause of FE can be omitted in the definition of 'explains'; the remainder of the discussion could then be construed as an account of \textit{intentional explaining}.
\item \textsuperscript{28} When I try to explain something and fail, the most natural way to describe the situation seems to me to be in the way I just did—my failure to convey the relevant information condemned me to merely trying to explain, rather than explaining. Thus, 'explains' is naturally thought of as a success term. Similarly, if a highly unlikely bout of Tourette’s syndrome caused me to unwillingly spout the words of an explanation, it seems most felicitous not to describe this event as one wherein I explained anything. Thus, 'explains' is naturally thought of as denoting intentional action. In any event, one who does not share these intuitions can just treat the present usage as a refinement for theoretical purposes, and can define other senses derivatively (see previous note).
\end{itemize}
purposes, though, any reader who doubts that 'explains' is typically used as a success term or as a term denoting intentional action can regard the above definition as a disambiguation—or at worst a refinement—for theoretical purposes.\textsuperscript{29}

Three other features of FE, however, go beyond choices about whether to treat explaining as necessarily successful or intentional; these features merit further attention here. First, FE treats the vehicle of explanations—that by which the explainer creates understanding—as representations. This treatment is motivated by two factors: first, the theory of understanding I favor (see §5), treats understanding as the possession of a particularly useful kind of representation. Therefore it is not implausible to regard the communication of representational content to be a typical cause of understanding. An even stronger motivation for taking the vehicle of explanations to be representations is that—at least intuitively—only representations seem to be the right kind of thing to be utilized in explaining acts. While there is an extensive literature on whether it is objects in the world or representations thereof that are the "real" explanations, everyone (in this debate) seems to acknowledge that representations are a necessary component for communicating explanations \textit{in explaining-acts}. The strategy of taking explaining-acts to be prior to explanations thus necessitates taking seriously the foundational role of representations. It is compatible with FE that the "real explanations" are still those things in the world picked out by the representations selected by FE—I do not know what would be at issue in such a dispute, or how to adjudicate it.

\textsuperscript{29} Similar remarks apply to the fact that only improved understanding is required—perhaps some readers will take what is really being defined to be 'partially explaining-act', but on most accounts of understanding there is no obvious threshold that would do as a measure of success for an explaining-act.
The second feature of FE that merits explicit defense is, regrettably, something of a lacuna in the view—the admittedly unsatisfying demand that explanations cause understanding "when properly internalized". Representations can cause improved understanding in all sorts of ways. Suppose, for example, one utters a magical incantation that causes the audience to suddenly develop all the brain structures upon which improved understanding of some proposition p supervene. This does not seem to be an act of explaining, yet it is difficult to articulate precisely what it lacks. Holding everything else constant—notably the intention of the explainer and the necessity that his representation have that very content in order to produce improved understanding—something still seems missing. The problem is a familiar one—in order to count as explaining, the act must cause the understanding, but it must do so in the right sort of way, without causal deviance. One might suppose, e.g., that for an expression to count as an explaining act, it must cause improved understanding in the normal way—that is, by the proper functioning of whatever mechanism or mechanisms in the audience's mind are generally responsible for improvement of understanding in standard operating conditions. Cognitive scientists make use of the idea of representations being "processed" without concern for deviant causation, so there is reason to be hopeful that the question will eventually be illuminated empirically; for now, it is probably best to remain neutral regarding the precise nature of proper internalization. Thus the phrasing in FE is

---

30 This is intended as a normative use of "normal", not a statistical use. For an account of the sort of normality pictured, see Millikan (1995). The term "normal" is hers, though "natural" might be a better description of the envisioned characteristc.

31 It has been suggested to me that I cannot remain entirely neutral regarding this, if I want to maintain that my account is genuinely reductive. There is some worry that, at least in other areas, what ends up counting as the "right way" for a particular state (e.g. the perception of redness) to be caused ends up
intended to remain as uncommitted as possible regarding the question of what the right way for the expression to cause the understanding is—so long as the representation expressed is internalized properly, however that may be, that is all that matters.

The third notable feature of FE is its apparent lack of any veridicality requirement. This should not be taken as an indication that, if FE is correct, any fiction that satisfies the curiosity of an audience counts as an explanation. The important constraint on the factiveness of explanations comes not from FE itself, but from the connected account of understanding. As will be discussed in §3 and §5, plausible theories of understanding demand that genuine understanding have a certain measure of veridicality. For example, on a view that treats understanding as a matter of ability, one only understands when one can successfully (and non-accidentally) produce certain results. Theories of understanding that lack any external check on what counts as genuine understanding will, for that very reason, fail to provide a smooth account of the fact that we commonly attribute to people (e.g. alchemists) a mistaken belief in their own understanding. Thus, FE does not by itself force one to accept (and, when paired with a plausible theory of understanding, likely forces one to reject) that anything that plays to audiences’ intellectual biases counts as explanatory.

being individuated by its causing of that very state. However, while I would prefer not to give up on producing a reductive account of explanation, the production of a non-reductive but illuminating account would hardly be a disaster.

32 A weaker guarantor of veridicality in explanations comes from the fact that explanations are always representations. While this is not the place to develop a theory of representation, it is plausible that in order to count as a representation at all R must be veridical in some sense.

33 In the opening paragraph of her piece on the relationship between understanding and the facts, Catherine Elgin (2007) notes that "pretty plainly understanding somehow answers to facts. The question is how." (33)
If the account sketched above is correct, it provides an elegant illumination of the failure of earlier attempts to give an account of explanations in terms of some particular formal structure or privileged contents. Most such accounts sought to provide intrinsically characterizing features of explanations; but on the current account explanations are characterized extrinsically by the role they play. The notable exceptions are earlier pragmatic accounts, which, while correctly focused on explanations' role, did not (I argue in §4.3) properly identify what that role is.

A further theoretical benefit of the current account is that, so long as everyday speakers have a general sense of when they do or do not understand something, they will be able to apply the term 'explanation' to the correct extension *without* requiring any knowledge of specialized content or structural forms. (Admittedly causal accounts have the same virtue, so long as people are generally able to identify causes. It is less clear how other accounts of explanation—particularly those accounts that take explanations to require some fairly sophisticated logical structure—can be as readily related to everyday practice.34) Given this, it is natural to wonder why the importance of understanding to characterizing explanation has been so long overlooked.

In fact, the usefulness of understanding for the purpose of characterizing explanation has not been *overlooked* at all—rather it has, from the inception of the

34 See §5.1 for a discussion of one strategy to relate D-N explanations to actual explaining acts.
subject in its present form, been explicitly denied. When Hempel & Oppenheim released their work on the logic of explanation, they articulated in painstaking detail very precise criteria for what qualified something as an explanation. Understanding, by contrast, was cast aside as either un-respectably phenomenological or a derivative effect of explanation (that thus requires a fully formed account of explanation before accounting for understanding):

But it is important to distinguish here understanding in the psychological sense of a feeling of empathic familiarity from understanding in the theoretical, or cognitive, sense of exhibiting the phenomenon to be explained as a special case of some general regularity. (Hempel & Oppenheim, 1948, p. 17)

The dilemma Hempel & Oppenheim present is telling: understanding is either merely a feeling, or else it is the derivation of the phenomenon from some general law. In the former case, understanding as "a feeling" is not of interest to cognition proper—in the latter case, it is simply what one knows when one possesses an explanation according to Hempel & Oppenheim's Deductive-Nomological model. They are clearly correct that in neither case could understanding play a valuable role in elucidating explanation.

In the next section, I argue that Hempel & Oppenheim's dilemma overlooks something—the possibility that we can characterize understanding independently of explanation yet in a way that is nevertheless not merely phenomenological.

---

35 It has been suggested that this reading is uncharitable, and that everything they say is compatible with the existence of a third use of 'understanding' that is neither phenomenological nor rooted in explanation. I do not want to commit to any particular exegesis, but it seems fair to suppose that the failure of Hempel & Oppenheim and many of their intellectual descendents to discuss any other sort of understanding suggests that they at least did not place much significance on such an alternative.
§3 The Respectability of Understanding

At first blush, one can sympathize with disregarding understanding as a legitimate basis for philosophical theorizing. Without a fully articulated account of understanding, one might suspect that it is a merely phenomenological state, the possession or absence of which depends entirely on a third-person-inaccessible feeling on the part of the understander. As recently as 2002, J.D. Trout argued that not only was understanding unhelpfully phenomenological, but actually so vulnerable to idiosyncratic prejudices and biases that it is inimical to scientific progress. To reiterate the dilemma attributed to Hempel & Oppenheim above: either understanding is characterized almost entirely in terms of its connection to explanation, or it is characterized entirely in terms of its phenomenology. In neither case can it do independent productive work in elucidating anything about explanation.

These choices are, however, not exhaustive; there might be an account of understanding that is neither steeped in phenomenology nor so closely tied to explanation as to be unhelpful. Quite recently, a variety of theories of understanding have emerged that do not require being analyzed in terms of explanation and that are not merely phenomenological in any sense that should be disquieting to the philosopher of science. Some theorists (e.g. Knuuttila & Merz 2009, tracing largely to Morgan & Morrison's 1999 *Models as Mediators*) contend that understanding involves possession of an appropriate model of the understood. More recently, a line of thought tracing to de Regt
& Dieks's 2005 "A Contextual Approach to Scientific Understanding" holds that one understands in virtue of possessing certain abilities to work with the understood.\textsuperscript{36} In epistemology (as opposed to the philosophy of science) understanding is often taken to critically depend on beliefs' coherence within an overall system of beliefs. This view steers clear of the merely phenomenological as well, in part by explicitly adding a veridicality requirement—proponents of such a view\textsuperscript{37} require that as a system the whole set of beliefs must be \textit{at least} "tethered"\textsuperscript{38} to the facts. All three types of views concede that discovery of the very same facts can cause understanding in one person but not another—to that extent, understanding is subject-sensitive. However, on none of the views is understanding itself subjective or phenomenological. (What counts as "improved quarterly earnings" depends a great deal on the subject to which it is applied—it requires a bit more funds for something to count as improved quarterly earnings for Apple than it does for me—yet there is nothing about the property of having improved quarterly earnings that is thereby rendered scientifically unacceptable.) Whether one possesses a model or an ability can be judged from a third-person perspective, and indeed can be misjudged from a first person perspective—to that extent, at least, such possession appear decidedly non-phenomenological. Admittedly different things can cause or fail to cause understanding in different people; however, this is no more a threat to the objectivity of

\textsuperscript{36} Admittedly, their view of understanding a theory (as opposed to a phenomenon) does invoke a Hempelian model of explanation. It would be a relatively straightforward project to excise this; even as is, since the explanatoriness of the explanations does not play into their view, one could simply describe the relevant features of the Hempelian model without ever invoking explanation.

\textsuperscript{37} E.g. Elgin (2007) or Kvanvig (2003).

\textsuperscript{38} The term "tethered" is from Elgin (2007, p. 35), who attributes it to Plato. She argues against the even stronger requirement that understanding be factive.
understanding than is the fact that different treatments work or fail to work on different people a threat to the objectivity of medicine.

For present purposes, it does not matter which of the accounts of understanding is correct—if it is possible to formulate some non-phenomenological account of understanding that is not defined wholly in terms of explanation, then it is worth reopening the possibility that what explanation is is fundamentally tied to its role in producing understanding. To the best of my knowledge, this upshot on the new work on understanding has not been drawn out by those in the area.

§4 Functional Explaining\(^{39}\) and the Philosophy of Scientific Explanation

Suppose one believes that, in any instance of speaker B explaining proposition p to audience A, B produces a representation that at least could cause A to better understand p, and that one further agrees to (at least for the moment) restrict one's usage of the term "explains" to intentional, successful acts. We knew all along that explaining should cause understanding, but one might think that this goes hardly any way toward differentiating explanations from other sorts of representations and/or propositions. There is a familiar litany of allegedly necessary features explanations need to possess: a deduction from a natural law of the explanandum, a unification of one's information by means of logical subsumption, an exhibition of a particular causal process or mechanism,

\(^{39}\) For the remainder of the chapter, I refer to the total view, consisting of UBEE and FE, merely as "FE". The three theses--UBE1, UBE2, and FE—can all be accepted or rejected independently of each other. However, when comparing my view to extant accounts of explanation, it is significantly easier to treat the theses as a package.
or some preferred pragmatic significance. (FE is an instance of this last schema, provided one understands "pragmatic significance" broadly enough. I believe it is largely unique though in requiring a particular cognitive effect of explaining, though, as will be discussed §4.3.2, Peter Achinstein's account comes close.) However, the requirement of any of these conditions faces well-known counterexamples. It remains to be shown both that their failings do not also afflict FE in some way and, just as importantly, that FE can capture those intuitions about cases that made the other conditions seem plausible in the first place. Both of these tasks (while not trivial) are easier than they sound, and for the same reason.

The key is that FE is such a radically different kind of constraint from the sort typically offered by earlier accounts. By slicing the conceptual pie in a very different way, FE can incorporate the positive examples taken as evidence for other views while having most of the alleged counterexamples fall to the wayside. It accomplishes this neat feat by simply taking the logic behind the pattern of positive examples and counterexamples more seriously than any local attempt to accommodate specific data-points. The counterexamples employed against old views either find objects the account in question dubs 'explanations' that intuitively do not seem explanatory or find objects the account fails to proclaim are explanations that intuitively do seem explanatory. What is it, however, that makes something "intuitively" seem explanatory or unexplanatory? One hypothesis is:
(H) Competent speakers, and philosophers in particular, can intuitively judge tolerably well what sorts of things would cause improved understanding in us, and are inclined to proclaim those things and only those things 'explanatory'.

If this hypothesis is correct, then not only will FE divide the cases into 'explanations' and 'non-explanations' in precisely the right way, but it will do so, in most cases, obviously. Counterexamples to older views that pointed out some object that the theory in question counted as an explanation but that did not "seem explanatory" (or, per H, seem likely to cause understanding) will not be counted by FE as explanations precisely because they did not "seem explanatory". It is important to note that H is an empirical hypothesis regarding how philosophers react to various counterexamples, rather than a conceptual claim about what either explanation or understanding is; we will thus be able to put it to test by applying it to examples. If H is true, then not only will FE correctly categorize negative counterexamples as non-explanations and positive counterexamples as genuine explanation, but it will usually do so quickly and intuitively.

In the remainder of this section FE will be compared to the main extant accounts of explanation—in general, FE will be immune to the standard counterexamples leveled against other accounts, exactly as predicted by H.

Before looking at the specifics, it is worth noting that, as FE defines explanation functionally, one would expect that different sorts of contents and structures could be

---

40 At this point nothing need be said about the basis of such judgments; specifically, even if we make such judgments on the basis of a particular phenomenological sensation, the phenomenology would only be serving a benign evidential role.
realizers of explanations. To that extent earlier accounts would be expected to have
gotten a large amount of the story right regarding what counts as an explanation, but
erred insofar as they claimed to provide the full story or the reason why some class of
things was explanatory.

§4.1 D-N & Unificationism

The most straightforward counterexamples to the explanatory necessity of
derivations of the explanandum from some law of nature are simple causal claims.
Michael Scriven (1962) points out that a caveman would have been in perfect position to
explain the presence of a stain on the carpet by citing the fact that my knee hit the desk
and knocked over the bottle of ink, well before the caveman could state the laws
governing force transfers and liquid diffusion.

One might object that Peter Railton (1980), in producing his D-N-P account, has
already explained away these alleged counterexamples by distinguishing between
particular explaining events and the "ideal explanatory text". Those proposing the
counterexamples equivocate between explaining acts, which sometimes lack all the
features of a real D-N explanation, and the real explanation to which they somehow
allude, which possess these features (or very similar ones). James Woodward (2003
chapters 4.2-4.3), however, persuasively argues against these "hidden structure"
strategies. Simplifying his nuanced argument, the central issue with such strategies is that
the precise relationship between acts of explanation and the ideal explanatory text is
either too mysterious to be helpful or, if articulated precisely, issues clearly problematic predictions. Moreover, there is something methodologically unsettling about making the claims of a theory so radically divorced from the "testable" cases of explanations as they actually occur in the world.

How does FE fare against the ink-spilling example? We would naturally take the caveman under discussion as gaining some improved understanding of the stain's presence from the putative explanation. Thus, FE would count the act as an explaining-act and the caveman would be counted (derivatively) as possessing an explanation. The case worked out well for FE, and, as expected, did so rather obviously.

FE can, however, make perfect sense of why DN struck many as a plausible view: in a host of cases, apprehending the link between a phenomenon and a law of nature can facilitate understanding. It seems safe to say that acquiring a representation of Newton's laws does commonly lead to greater understanding of a wide variety of the specific phenomena subsumed by them. Thus, the extension D-N assigned to "explanation" had non-trivial overlap with the corresponding extension according to FE—it got many cases right. Moreover, if one really could account for all cases of explanation by appealing to a particular structure found in a few paradigmatic examples, it would be quite theoretically appealing to do so—in fact, FE can account for this appeal uncommonly well, by noting the improved understanding one typically gets via appreciation of the common structural features of some relevant class. I understand skyscrapers better when I learn a bit about the properties of steel skeletons. However, declaring that all and only things with the structural features cited count as explanations overgeneralizes by ignoring the functional
nature of the original target—a definition of skyscrapers in terms of some favored construction would rather miss the point. If such a definition is appealing, then it will doubtless be right much of the time. In fact, it could, as a matter of empirical accident, even be extensionally adequate all the time—it would still rather miss the critical relational property of skyscrapers (relative height), fail to generalize to future and counterfactual cases (titanium-arch based skyscrapers), and generally bear little correlation to the reasons we care about them (possibly aesthetic?).

There are of course many variants of D-N, but anything that is based on a foundation of explanation-as-derivation will run into the same basic problem in that our actual explanations frequently do not have this structure. Nor would it do to consider this deductive structure a regulative ideal, as it does not intuitively seem (to me at least, nor I gather to Woodward) that counterexample explanations such as the Scriven's ink-stain would be the slightest bit improved by spelling out a precise derivation of the existence of the stain. Given the computational complexity of such a derivation, one might argue that such a spelling out would make the explanation substantially worse.

Though they are perhaps less problematic for D-N, the historical importance of two other classes of counterexamples makes them worth noting—explanations with irrelevant information and explanations that deduce the occurrence of a cause from an effect. We can derive from the fact that a man takes birth control, along with various biological laws, that he will not get pregnant; we can also derive from the length of a shadow and the angle of the sun, along with various physical laws, the height of the flagpole. However, despite their adherence to D-N form, we do not take such arguments
to have explained either the man’s failure to be pregnant or the height of the flagpole. Intuitively, FE gets both the male birth control and the flag pole cases right—someone who receives an "explanation" of barren men in terms of birth control or flag pole heights in terms of shadow lengths positively misunderstands the phenomenon explained. Importantly, the fact that understanding does not reduce to mere phenomenology means that we can rightly say recipients of male birth control and shadow-to-flag-pole explanations misunderstand even if they would emphatically claim that they did understand. (For a discussion of what justifies the claim that they do not understand, see §4.4 and §5.)

The most important direct successors to D-N are the unificationist accounts of Friedman and Kitcher. Both accounts accept that an explanation is some type of argument whose conclusion is a statement of the explanandum. They differ from D-N and each other only in how they pick the explanatory arguments from the unexplanatory—where D-N requires natural laws, unification requires an explanatory argument to enable us to better systematize the knowledge we possess. These views, however, do not obviously fare better with respect to the simple causal explanations that plagued D-N. First, it is not clear such causal stories must always take the form of arguments to be explanatory. To the extent that we countenance non-arguments as explanations, the unificationist must either be a revisionist with respect to our practice, or appeal to the same sort of hidden-structure strategy that appeared inadequate with respect to D-N. Second, successful explanations do not always unify. I might be able to explain the presence of the ink-stain by informing the audience of its causal history even if the audience has no particular
beliefs about ink or knee/desk collisions.\textsuperscript{41} As before, FE handles such cases easily. Moreover, unificationist accounts sometimes bring along costs all their own, such as the radical irrealism Kitcher's view forces on its adherents (see chapter 8 of his 1989).

The appeal of unificationism is the same as that of D-N—Newton's laws, in addition to all else they do, unify our store of beliefs (per Friedman) and enable for more generally applicable "argument patterns" (per Kitcher). Hence those laws get counted as positive evidence for unificationist theories, though, from the perspective of this essay, unificationism fails to capture what is so special about those laws (i.e. that knowing them is generally conducive to understanding).

§4.2 Causal Accounts

According to causal accounts, one explains an event if and only if one cites its cause, and an explanation is either that citation or the cause itself. When the operative notion of "cause" is taken to be dependence in some relevantly interesting class of counterfactuals, a la Woodward, the view correctly classifies many cases within a large swath of empirical sciences (which fact would explain its appeal). However, there are still good reasons not to demand that all explanations be causal. Most importantly, there seem to be explanations given regarding domains where talk of causality (even broadly construed) simply does not seem to make sense—explanations in mathematics are the most obvious.

\textsuperscript{41} The committed unificationist does have some possible responses that regrettably cannot be explored for space constraints. For a more thorough discussion and rebuttal, see Woodward (2003, Chapter 8).
example, but explanations in any subject matter that crucially involve necessary truths will all be unaccounted for. Where there can be no counterfactual variation, there can be no interesting counterfactual dependence relations. Woodward acknowledges this, and restricts his scope to scientific explanation—at one point he even claims to only be elucidating specifically causal explanation only (Woodward 2003, Chapter 1.1).

At least as pertains to the weaker restriction to scientific explanation, however, it is not clear that Woodward can bracket concerns about mathematical explanation so easily. While it is plausible that explanation would be radically different in domains as diverse as science and mathematics, it would be difficult to neatly segregate those explanations within the sciences that crucially turn on some mathematical demonstration from those that do not. If I wanted to explain to an evolutionary biologist why honeycombs are made of hexagonal compartments, the key would likely be demonstrating to her the unique efficiency of hexagons to segment a plane. This, however, is a mathematical, necessary, and invariable fact. There will be non-mathematical facts invoked in the explanation as well, such as that natural selection will tend to favor bees who use their resources more efficiently, but in this context, those do not seem to be the facts that are doing the explanatory work. We can suppose that the only part of the explanation that was not already familiar to the scientist was the mathematical element, which brings into focus the fact that a necessary truth is the lynchpin of the explanation in this context.

FE correctly predicts that the mathematics is genuinely explanatory in this context and, true to the pattern, does so rather easily. The evolutionary biologist already
understands natural selection fairly well, and so representations geared towards improving her understanding of that directly are bound to flounder. However, she might not understand one bit of planar geometry, and so that aspect of the explanation might genuinely increase her understanding of the relevant phenomenon.

All of this is compatible with agreeing with Woodward that he has given a perfectly good account of causal explanation, which at one point is all he claims to be doing. But, if so, it is reasonable to ask what more there is to scientific explanation, let alone explanation in non-scientific domains. Moreover, if our question is whether Woodward has found a necessary constraint on the content of explanations, the stronger restriction of his account to just causal explanations renders that question moot. Clearly it is a necessary condition for something to be a causal explanation that it say something about causation, but it is just as necessary that a mechanistic explanation say something about mechanisms and a chair-y explanation say something about chairs. There is no interesting lesson to be drawn about explanation generally from these observations.

While there are non-counterfactual accounts of causal explanation, they are typically more restrictive than Woodward's counterfactual account, not less. None would countenance mathematical explanations as causal. For example, no fixed quantity is transferred in the course of a mathematical explanation, and so it would not count as causal on Wesley Salmon's most recent view (see a summary in his 1989). Woodward's account is the most liberal in that it (or a variant of it), uniquely, might able to count compositional facts as causes, allowing for example explaining this statue's melting being
32° F by pointing out that it is made of ice. If even Woodward's criteria are not necessary, then no other causal account will fare better. FE does.

§4.3 Pragmatic Accounts

If any account that specifies what counts as explanations in terms of candidate explanations' role counts as a pragmatic account, then FE is included, and so there is no way to mount a defense of FE by decrying pragmatic accounts in general. We can, however, compare FE to the two main alternative pragmatic accounts on offer.

§4.3.1 Van Fraassen

Bas Van Fraassen's pragmatic account of explanation has been accused of being loose; Kitcher and Salmon (1987) argue that, for all the constraints Van Frassen has placed on explanation, any fact could still explain any other proposition in some context or other. It might thus strike some as odd to say that even Van Fraassen's constraints are in some way overly restrictive. However, while it is true that any proposition could explain any other in some context or other, it is still an open question whether Van Frassen's account restricts certain propositions from being explanatory in specific contexts where they intuitively should be. I believe that it does. On his account, one explains P by producing some A that bears a contextually determined relevance relation R to P and that probabilistically favors P over the other members of some contextually
determined contrast class X (Van Fraassen 1980, p. 147-50). While it has not drawn as much attention as other aspects of the account, the requirement of probabilistic favoring is a bit surprising, as there is reason to doubt that all explanations favor their explanandum. For example, while wearing seat belts greatly reduces the risk of dying in a car accident, there is also a small percentage of car accidents where someone lives precisely because she did not have her seat belt on. If the driver is flung from a car which then promptly explodes, one might ask why she lived. The explanation could quite possibly be that she lived (as opposed to the other member of the contrast class {lived, died}) because she was not wearing her seat belt, even if this actually lowered the antecedent probability that she would make it out alive (we can suppose that the explosion was an improbable result of a genuinely indeterministic process). FE, by contrast, seems to get this case right, as we intuitively do understand her survival better when we know that she was not wearing her seat belt.

If one removes the probabilistic favoring aspect from Van Fraassen's account, one is left with only the requirement that the explanans bear some contextually determined relevance relation R to the explanandum, but not to the other members of the contrast class. This stripped-down account is not too restrictive, but it also does not add any constraint in addition to FE. FE merely specifies one relevance relation (the "causes understanding of" relation) as relevant in all contexts, which possibility is not ruled out by anything Van Frassen says.

Thus, we are left with the choice of either taking Van Fraassen's requirement of probabilistic favoring as a critical part of his theory or not. If we do, then his theory is
inadequate. If we do not, then his theory requires nothing of explanations not already required by FE. Either way, there is no reason to look to Van Frassen for additional constraints.

§4.3.2 Achinstein

Another well known pragmatic account of explanation is due to Peter Achinstein, whose account is in some ways similar to FE. Achinstein says:

[Person] S explains [proposition] q by uttering u if and only if S utters u with the intention that his utterance render q understandable by producing the knowledge, of the proposition expressed by u, that it is a correct answer to Q. [Q is the direct form of a question whose indirect form is q. For example, if the proposition being explained is that Nero fiddled, Q is "why did Nero fiddle?"]

—Achinstein 1983, p.72 (definition of Q from p. 17—there are some additional constraints on the structure of Q that are not relevant for present purposes)

Achinstein thus accounts for the nature of explanation in terms of explaining, which is in turn reduced to its connection to understanding—resulting in an account that is structurally very similar to FE. What is historically striking, however, is how little Achinstein's work has impacted the trajectory of the philosophy of explanation. While his rigor is generally admired, few in this area grapple with his positive theory. If the present essay contributes nothing more than to resurrect a promising research trajectory that has lain largely dormant for thirty years, I think that would mark a valuable contribution.
Broad similarities aside, the details are important, and therein can be found some salient differences. Specifically, there are three main distinctions between Achinstein's view and FE.

The first difference is that, rather than focus on understanding itself, Achinstein's account instead characterizes explaining solely in terms of the intention of the explainer. This construction, however, severs the direct link between explaining and understanding and allows for some bizarre acts to count as explaining, if the speaker has strange enough beliefs regarding what sorts of things constitute understanding. This disconnect also makes our quick ability to intuitively judge whether something is genuinely explanatory a bit more puzzling—it is less obvious how we could have a sense that tracks when other people possess a particular intention than how we could have a sense of when we are in a particular mental state.

The second difference is that Achinstein (23) weds understanding, and hence explaining, very tightly with knowledge of propositions. However, if we are to take scientists' word regarding what constitutes understanding, this is overly constraining. Though neither of the following claims is uncontroversial, direct study of scientific practice suggests that, for example, understanding in economics (Boumans 2009) and biology (Leonelli 2009) can come from model manipulation. Moreover, the benefits of diagramatic reasoning to understanding in mathematics (see the discussion in Avigad 2010, p. 15-17) and cognitive science (Bechtel 2008, p. 19-20) become difficult to reconstruct if one restricts understanding to a type of knowledge of propositions.
The third and most important difference between Achinstein's view and FE is how the views can be extended to a theory of explanation assessment. The natural way to evaluate an explanation according to FE is by the extent of the improved understanding it can be utilized to provide. How this evaluative theory is implemented will depend on the precise theory of understanding—on an understanding as model-possession theory, the quality of an explanation will depend on the extent to which it enables the audience to construct better models of the explanandum. Any veridicality requirement will thus be a result of the theory of understanding (in this case, the fact that the quality of a model depends to some extent on possessing a certain degree of fit with the world). Explanation evaluation for Achinstein, however, can be a bit trickier. Veridicality is demanded as an explicit evaluative criterion—the quality of the explanation depends in part on the truth of the proposition being used to explain. However, the sorts of propositions Achinstein considers include claims such as "the explanation of why that occurred is that p" (71), the truth of which is the sort of issue a theory of explanation is supposed to provide. Achinstein (71-72) addresses a related circularity worry, claiming that the criteria for evaluating explanations in fact do not make appeal to explanations themselves—however, if truth respects disquotation, then requiring that "the explanation of why that occurred is that p" be true in turn requires that p be the explanation of why that occurred, potentially reintroducing the feared circularity.
§4.4 A Response to an Objection

In order to remain neutral regarding which is the proper account of understanding, the previous sub-sections relied on our general intuitions regarding which cases provide understanding and which do not. This generality, however, opens the door to the objection that the state of mind most likely to be guiding these intuitions is the phenomenological sensation typically associated with understanding. But it was argued in §4 that genuine understanding is best understood non-phenomenologically. The intent of §4.1-§4.3 was to show that those acts which cause understanding are explaining-acts, but all that was demonstrated was that those acts that seem to cause understanding seem to be explaining-acts.

The objection, however, turns on an equivocation regarding what it is for something to seem to cause understanding. On the one hand, there is the sense of understanding—associated with a particular phenomenology of "getting" something—that would be questionable as a guide to assessing understanding claims. If this section merely demonstrated that that explaining-acts seem to cause understanding in that sense of "seem", then it would be left to show that we have any reason to believe they actually cause understanding. That is not, however, the pattern that was exhibited. What was shown was not that explaining-acts seem to cause understanding in the sense that they produce a particular phenomenology, but rather in the sense that from a third-person theoretical perspective we judge them to be cases where understanding was likely to be
caused. The fallibility of the sense of understanding is no more grounds to reject judgments of understanding-cases as a basis for evaluating a theory of explanation than is the fallibility of individual conscience grounds to reject judgments of ethical-cases as a basis for evaluating a theory of normative ethics. The reliability of our judgments pertaining to whether a described scenario is an explaining-act is even more secure than our judgments of whether understanding would be produced, as there is no particular phenomenological association with observing an explaining-act to potentially mislead us.

§5 Understanding

The previous section showed that our judgments regarding classic examples of explaining conform to a particular pattern—we only count those acts that produce understanding as genuinely explanatory. It would be natural to wonder, however, what a plausible account of understanding might look like and whether it would actually predict the right results with regard to these cases. The most important point to note here, however, is that the construction of an account of understanding is an independent project. For all that has been said, understanding could even be a primitive epistemic notion—so long as it is neither analyzed in terms of subjective phenomenology inimical to scientific research nor in terms of explanation, any old notion of understanding will do.

As it happens, however, we can do better in articulating a notion of understanding that is both independently plausible and predicts the right results for evaluating cases of putative explaining-acts. Producing and defending a full account of understanding is
beyond the scope of this chapter, though it is a project I undertake in detail in chapter 5. It might, however, be illuminating to take just the product of that other inquiry as given,\textsuperscript{42} and see whether it correctly classifies some of the examples of putative explaining-acts from the previous section. It does; such success argues for the correctness of FE (and provides some small evidence in favor of the account of understanding I favor).

Understanding is, I claim, a manner of representing that which is understood in a way that is useful in the relevant context. More precisely:

URM (Understanding as Representation Manipulability): A statement, attributed in context C, that thinker T understands object o, is true if and only if T possesses a mental representation R of o that T could (in counterfactuals salient in C) modify in small ways to produce R', where R' is a representation of o and possession of R' enables efficacious (according to standards relevant in C) inferences pertaining to, or manipulations, of o.

This accounts satisfies the desideratum of neither relying on a particular phenomenological sensation nor in any obvious way reducing understanding to explaining. Moreover, it does, as hoped, make the correct predictions with respect to paradigm cases of explaining and non-explaining. While a full treatment of each example is impossible in this space, one example of how a successful explaining-act causes understanding and one example of how a failed explaining-act does not will hopefully suffice to illustrate how other examples work out.

\textsuperscript{42} To be clear, neither the correctness nor general applicability of the account are supposed to be at all uncontroversial. This section is simply focused on how such an account, if it were broadly acceptable, could be usefully deployed.
Coming to understand the presence of an ink stain by being told of its causal history creates a representation that enables all sorts of (typically) valuable inferences, e.g. pertaining to how it could have been prevented, what sort of cleaner might best get it out, and what an ink stain on the other side of the desk might look like. The judgment that the recipient a successful explaining-act invoking this history would gain understanding is thus validated. Note that there are contexts—such as, perhaps, a physics laboratory—where representing the causal history of the stain would not enable the sorts of inferences we would count as relevant. But it also seems right to say that, in those particular contexts, the caveman-knee-hitting story would not suffice to produce understanding.

Contrast the ink-stain case with a paradigmatic pseudo-explanation, such as astrological explanation. There are no doubt many people who, upon learning that former romantic partners have conflicting zodiac signs, get the sensation of understanding their relationships' failures better. Such people would likely, as predicted by the empirical hypothesis connecting our sense of understanding to our sense of how explanatory something is, be prone to saying that the astrological explanation is genuinely explanatory. They would be mistaken. But they would also be mistaken if they believed that they understand the target phenomenon any better than they had before. Though they possess the right feeling, the inferences they might draw would be unreliable—contra their beliefs, had their partners been born under different signs, the relationships would still likely have been failures. Even if they chanced to live in a world where inferences based on astrological facts happened to be reliable, the invocation of counterfactuals in
URM demands that the inferences would be useful blocks their claims to understanding (in the vast majority of contexts).

Since the successful explanation being mangled is in both cases a causal one, one can see why the defunct explanation fails to cause improved understanding—in contexts where the audience is concerned with the causal history of the explanandum, the representations presented in the non-explanations fail to be reliable in the right sort of way. Similar points can be made about the famed cases of irrelevant information and causal asymmetry. In any context, knowing about the man’s taking birth control does not enable me to make any new inferences regarding his failure to be pregnant, nor does it enable me to better control future cases of male non-pregnancy. Similarly, knowing the length of the shadow does not enable me to make the sorts of inferences or exert the sort of control relevant in most contexts in which we might consider the height of the flag pole. Arguably, Van Fraassen’s (1980, p. 132-34) tale of a tower being constructed for the sake of casting a particular shadow might be an exception.43

Success with regard to a few examples of course does not prove that the point that genuine explanation is that which is conducive to genuine understanding, but it provides some evidence for such a relationship being that which underlies the pattern of explanatory successes and failures that beset other accounts. It is thus reasonable to shift the burden to the opponent of FE to find a suitable counterexample.

43 I deal with examples of explanatory asymmetry at greater length in chapter 7.
§6 Conclusion

In this chapter, I have suggested a new approach to the philosophy of explanation suggesting that we should produce an account first of explaining, and eventually of explanation, out of understanding. I accounted for the absence of this view in the historical literature by suggesting that modern accounts of understanding have only recently enabled this new approach. I argue that such an account is largely immune to the sorts of counterexamples that have plagued the philosophy of explanation for decades. Perhaps more importantly, it returns explanation from the domain of formalistic analysis to the everyday environs in which it belongs.
Chapter 4: A Review of Understanding

It has become evident that understanding, whatever that is, plays a critical role in determining what counts as an explanation and in how explanations operate. FE cashes out explanations in terms of explaining-acts, and explaining-acts in terms of their causal role in producing understanding. Absent at least some idea of what understanding is then, we will have very little to say about the nature of explanation. In this chapter I will look at theories of understanding present in the literature.

The sort of understanding of interest for analyzing explanations and explaining-acts primarily takes as its object theories, propositions, and material objects—not primarily words or sentences. While there has been much discussion of what it is for someone to understand the meaning of a word or sentence, words and sentences are far from all we talk of people understanding. Though the types of understanding are related, what it is to understand 'quantum mechanics' is very different from what it is to understand quantum mechanics. It is in this latter guise that understanding most obviously relates to explaining. We can educate someone to the point that they understand the meaning of 'quantum mechanics', but it is not clear how such education relates to explanation; by contrast, getting someone to the point that they understand quantum mechanics is often a clearly explanatory endeavor. I thus put aside questions of understanding meaning as a project to be undertaken elsewhere. The concept that would
be of most use in filling out our account of explanation is perhaps best characterized as *cognitive understanding.*

Unfortunately, the history of philosophical inquiry into understanding has been checkered, and it is difficult to draw out one cohesive narrative. However, we can see philosophers continually grappling with a series of related questions regarding the role of *understanding* in relation to other concepts. Such questions are:

1. What is the connection between understanding and explanation?
2. What is the connection between understanding and (explicit) knowledge?
3. What is the connection between understanding and ability?
4. What is the connection between understanding and a certain familiar qualitative "eureka" experience?
5. What is the connection between understanding and representation?

While at least the first four of these questions were always in the background of philosophical work on understanding, each of the first three received particular attention in its own historical movement. The first such movement consists of sociologists and continental philosophers who inquired into the empathetic and hermeneutic basis of

---

44 That being said, I will continue to just use the broader term 'understanding' for ease of exposition.
45 Linda Zabzebski, in her 2001 "Recovering Understanding", argues that while discussion of understanding dates back at least as far as Plato it tends to get crowded out of epistemology when concerns about skepticism refocus epistemologists' attention on the nature of justification.
46 For a thorough exploration of some of the more recent literature in epistemology, see Kvanvig (forthcoming).
historical understanding, as distinct from scientific explanation. These scholars were largely interested in how understanding of human behavior fit in to a world of explanatory natural laws. The second and most influential epoch began in the mid-20th century with philosophers' new found keen interest in the form of scientific explanation. Philosophers in this tradition assumed understanding was (at least typically) the product of good explanation, and so beliefs regarding the nature of understanding can be inferred from what product would be produced by an explanation according to the account. Since, according to the first such account of explanation (due to Hempel and Oppenheim), explanations were arguments with particular linguistic content, understanding came to be seen as possession of a certain type of propositional knowledge. This emphasis is still present in much of the literature today, with much of the literature devoted to trying to identify particularly explanatory knowledge, typically as a function of some privileged content (e.g. causal, mechanical, constitutive). While the connection between understanding and ability has been perceived as important at least since Wittgenstein's discussion of understanding in the *Philosophical Investigations*, there has been a resurgence recently of systematic investigation into this particular connection. This resurgence began in "A Contextual Approach to Scientific Understanding" by De Regt & Dieks in 2005, then was vastly amplified and altered in the 2009 anthology *Scientific Understanding: A Philosophical Perspective*. This thread focused on understanding as seen in practicing sciences, which resulted in a focus primarily on the connection between understanding and certain practical abilities.
The particular qualitative experience of "eureka" moments is interesting: while no theorist of understanding takes it as a necessary or sufficient condition of understanding in general, it is perhaps the most readily accessible feature of the phenomenon, and so merits some inquiry. Moreover, there is one domain of understanding wherein it is thought that a particular phenomenology might be essential—that is understanding of one's own thoughts. This link has been developed in the literature on the epistemology of thought.

The link between understanding and representation generally is the least well explored of the five. However, since on my own account understanding and representation are very tightly linked, this is an important connection to explore.

In the next five sections I will look at the each of these questions in turn. In §6 I will look at a suggestion by Jay Rosenberg that all there is to understanding is attributions of understanding rather than some underlying basis of attribution.

§1 Understanding and Explanation

In most of the recent literature on scientific explanation it is taken for granted that good and correct explanations at least typically cause understanding. The relationship explored in the continental literature is more complex. There can be found a well-known distinction between *erklären*, the kind of scientific explanation we engage in when accounting for the natural world, and *verstehen*, the kind of understanding we get of others' actions (writ large in the cases of history and art). Though human actions are part of the natural world, they were thought to require very different methods of study and
analysis. The thought expressed by such sociologists and historians as Weber appears to have been two-fold. First, the world is infinitely complicated, so we cannot possibly make sense of it as it is. Given that, we would expect to have a better grasp of others' actions then we would the natural environment—whereas in the latter case we have no principled means of dividing up the world, in the former we have some reason to hope that other agents will have constructed similar worlds to our own.47 We can understand others, whereas we can only explain nature, because only in the former case do we possess the invaluable tool of empathy.48

Putting the details aside, this bifurcation principally rests on an optimism regarding our epistemic access to human behavior as opposed to everything else. Even if one grants that we have privileged access to our own mental states, the optimistic generalization to the mental states of others is (absent further argument) a reckless inference from an existential generalization to a universal one. And even if one could overcome this hurdle, one must then face the fact that we always see agents acting in the same empirical context or other. In order to understand that a man is afraid, I need to (for example) have already understood that there is a bear coming at him, and in order to understand that the man's subsequent actions will lead to the bear's failure, I need to have

47 For an example of the sort of work in question, see Weber (1922/1978, p.7)—there he says “that which is intelligible or understandable about [an object] is thus its relationship to human action in its role either of means or ends…only in terms of such categories is it possible to understand objects of this kind.”

48 This section greatly over-simplifies an array of different views within the tradition broadly identified as continental. The principled reason for this is that much of the opposition to the dichotomy between erklären and versethen coalesced in the logical positivist school, and thus can be viewed as a precursor to the Hempelian line presented below (Hempel being of course a noted logical positivist). The less principled reason is that much of the literature is fairly impenetrable (to me), steeped in sociological and anthropological presuppositions and terminology.
already understood that the fearful man will choose to run—the two types of understanding are far too intertwined to be readily pulled apart. Von Wright (among others) accounts for this fact by proposing a sort of (temporal) zig-zag\textsuperscript{49} grasping of even the most ordinary historical circumstances—first we explain the empirical consequences, then we understand their effect on agents, then we explain the effect of the agents' actions on the empirical circumstances, then we understand the effects of the new empirical circumstances on the agents... This sort of answer does not mesh well with at least my own phenomenology when explaining/understanding agents in the world—it certainly does not feel to me that I am constantly switching back and forth between two totally different epistemic approaches. It feels like I am doing one thing—trying to understand people and their environment (as opposed to alternatively trying to understand people and trying to understand their environment). My phenomenology is not decisive (though one would think it would have more weight than normal on a picture that emphasizes our primary access to our own mental lives), but it does provide some reason to favor an account that treats people and the environment as roughly on a par as far as being objects of understanding.

Returning to the more recent analytical tradition, talk of understanding as a psychological phenomenon or even as a mental state fell into disrepute. Carl Hempel claimed that such a state of understanding had no place in a respectable philosophy of science (see n. 3, Hempel 1965 p. 413), but then in a famous passage cleared logical

\textsuperscript{49}This is literal—Von Wright's diagrams of the process (1971, p. 137, 138, and especially 143) are zig-zagged.
space for a very different type of understanding: by displaying in argument form the nomic predictability of the explained phenomenon, a DN explanation "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, 337 italics in original). If Hempel's view of understanding could be read simply off this passage, it would be that one understands a phenomenon when one could have expected it, and anything else that goes under the name "understanding" does not merit scientific consideration. Many commentators have pointed out that this is false—one could possess the most reliable barometer in the world yet still utterly fail to understand storm-fronts—yet what strikes me is that, taken on its own, the fascinating thing about this account of understanding is that it is obviously false, and yet absolutely no effort is made to argue for it. Hempel did not find his seemingly radical claim about the nature of understanding worthy of significant comment—real understanding just had to be whatever one came to possess upon being properly exposed to the appropriate explanation. Questions of barometers and storms were indeed pressing for Hempel, but they were pressing in forming his account of explanation—once that was settled, such obvious counterexamples to his account of understanding were most likely expected to go by the wayside. An account of understanding was, to Hempel, merely a corollary of the correct account of explanation. Whatever is produced by explanation, for Hempel, simply is understanding. Unlike the continental account, where understanding and explanation were not related at all (except to Von Wright, for whom they were related

50See e.g. Kim (1992)
only contingently), to Hempel what counted as understanding was fully constrained by what counted as explanation. There was no need for a separate account of understanding.

While Hempel was still in the process of articulating his views, there was already one philosopher who argued that he had the order of explanation backward. In his 1962 “Explanation, Prediction and Laws” Michael Scriven claimed that while explanations and understanding are constitutively linked, it is an independently correct account of the latter that fully constrains the former. If Scriven is correct, explanation is not particularly helpful when formulating an account of understanding; since explanations are whatever facilitate understanding in the right sort of way, there is little we can learn about understanding from looking at explanation.51

Most accounts have fallen somewhere between the extremes taken up by Hempel and Scriven, taking accounts of explanation and understanding as needing to be both independently plausible and mutually constraining. For example, while Michael Friedman argues that explanations on his account produce understanding, he does not merely identify understanding with that which is produced by explanation. Rather, he attempts to defend a particular account of explanation, a particular account of understanding, and the further claim that what is produced according to the further at least tends to create the mental state described by the latter. Accounts that posit this sort of connection between explanation and understanding face an explanatory burden not

---

51Scriven does not explicitly give an account of understanding. At times he speaks as if all understanding is causal, but I think his real view is expressed when he talks about the point of explanation being to demonstrate a connection between “the object of inquiry and the realm of [what is understood] in some comprehensible and appropriate way” (71). This puts him in broadly the same camp as Toulmin, only instead of reduction to the familiar Scriven emphasizes a connection to the already-understood.
faced by any of the other possibilities. Where the Hempelian and the Scrivenian needs only put forward only one independent account and a defense of the claim that the other is a parasitic phenomenon (understanding in the former case and explanation in the latter) and the continental philosopher needs put forward two independent accounts but no defense of any connection, the follower of Friedman needs two independent accounts and a further defense of their connection. This added explanatory burden leaves modern accounts of explanation and understanding (which largely follow Friedman's lead) more open to the sorts of counterexamples discussed in chapter two.

The most important point for present purposes is that, Weberians aside, most writers on explanation would happily concede some version of the claim that successful, correct explanations tend to result in understanding. This is important because there has been much more work done in the past 60 years on building precise accounts of explanation than there has been in building equally rigorous accounts of understanding. However, by looking at the products of extant accounts of explanation we will be able to explore a great many different views of understanding that have not been directly articulated.
Table 3 Explanation and Understanding--The Logical Space

<table>
<thead>
<tr>
<th>Necessarily Connected</th>
<th>Explanation Prior</th>
<th>Understanding Prior</th>
<th>Autonomous—neither prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hempel</td>
<td>Scriven</td>
<td>Friedman</td>
<td></td>
</tr>
<tr>
<td>Chart</td>
<td>Kitcher</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achinstein</td>
<td>Salmon</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wilkenfeld</td>
<td>Woodward</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Necessarily Connected</td>
<td>X&lt;sup&gt;53&lt;/sup&gt;</td>
<td>X</td>
<td>Weber</td>
</tr>
</tbody>
</table>

§2 Understanding and (Explicit) Knowledge

By far the most complex question in the literature on understanding pertains to the relationship between understanding and (explicit) knowledge. The question of whether understanding is a species of propositional knowledge is vexed, in large part by the fact that we are still not what the best account of propositional knowledge is. As issues regarding understanding and particular accounts of knowledge might be of independent

---

<sup>52</sup>In principle there is no reason that the logical space has to be carved up in such a discrete way. There is an infinitude of positions that maintain that explanation and understanding are constitutively linked in a complex web wherein each partially defines the other.

<sup>53</sup>That the bottom-left two squares are unoccupied is expectable, since if the two concepts are not necessarily linked it would not be clear how to even make sense of the idea that either is (explanatorily or ontologically) prior to the other.
interest to epistemologists, I will this put them aside and reexamine some of the possibilities in Chapter 6 (after my account of understanding has been place). In this chapter I will examine accounts that suggest connections between understanding and knowledge that do not depend on a particular epistemological account—views that identify understanding with knowledge pragmatically, such that one understands if and only if one knows the answer to a why question, and views that identify understanding not with any particular piece of knowledge but with global properties of a knowledge store.

§2.1 Understanding as Knowledge of an Answer

Declan Smithies (in conversation) suggested that a particularly common-sense notion of understanding would be that understanding is knowledge one possesses that answers some question relevant in that context. While there are several closely analogous views in the literature, all of them, upon closer inspection, prove importantly different from the Smithies proposal. I will discuss three such views—that of Sylvain Bromberger, that of Bas Van Fraassen, and that of Hasok Chang. While the first two are billed as views according to which what counts as explanatory (and hence, perhaps, understanding-producing) is a matter of presenting the answer to an appropriate question, when examining the formal details the importance of the question largely drops out. On

---

54 This view has the benefit of accounting simply for the apparent context-sensitivity in our understanding attributions.
Chang's account questions do play a critical role, but I argue that there is no way to articulate which questions are such that possessing the answers to them constitutes understanding.

The first question-based view of explanation in the contemporary literature was Sylvain Bromberger's 1966 account of explanations in "Why Questions". Bromberger contended that an explanation is called for when an event fails to conform to a perceived general pattern. For example, if it is generally the case that most compounds contract when they freeze, it is reasonable to ask why water expands, but not why lava contracts. The explanation takes the form of an ammendation to the perceived general law—typically the "unless" clause in a statement like "compounds contract when frozen unless they're composed of..." However, what counts as an explanation on this view is based on whether the general law being amended has the form of a general law and whether the amendment being offered has the form of an appropriate amendment. It is not in virtue of some statement's answering a question that it counts as explanatory; rather, it is in virtue of the fact that it completes an otherwise imperfect law that a statement both counts as explanatory and answers a question that is likely to arise. The account, for all its pragmatic trappings, is in the final analysis largely formal. But formal accounts clearly demand far more than the typical understander/explainer possesses—a grasp of first order logic is not a prerequisite to possessing understanding. (Perhaps a handful of important concepts from logical theory, like consequence, are required, but it would be far from the full corpus Bromberger needs for such sophisticated notions as the complements of abnormal natural laws.) In the case of explanation, it is traditional to declare that the
formal account gives the ideal explanation, to which actual cognizers only aspire to approximate. Whatever merits this idealization has in defense of an account of explanation, it will not do in elucidating understanding. Understanding, whether a mental state or an ability or something else, is a property of actual people—an account of ideal understanding makes no more sense than an account of ideal believing or ideal fearing.

Pragmatics play a more central role in Baas Van Frassen's account of explanation, according to which one explains by producing an answer that probabilistically favors the occurrence of the explanandum relative to some pragmatically determined contrast class. The requirements of probabilistic favoring and contrast classes restrict the account in ways would not work well as an account of understanding. Understanding how a mechanism or process works, for example cellular meiosis, seems to be a paradigmatic type of understanding, but it is difficult to fit into Van Frassen's framework. What is the contrast class? What does the content of our understanding suggest is relatively more likely than if what we understood was not the case? One could jury-rig answers to these questions, saying that the contrast class is the set \{Divides properly, Does not divide properly\}, and that the understood mechanism being in place favors proper division over a lack of proper division, but to seek out contrast classes and probabilistic relations in this way is to render the formalities of Van Frassen's account wholly inert. While Van Frassen focuses on the importance of questions in contexts, his view is overly restricted to a

---

55 On a careful presentation of this move, see Railton (1980), and for a rebuttal of its success see Woodward (2003, Chapter 4.3-4.4), summarized in Chapter 2.
particular type of question and a particular type of answer to serve as a likely account of understanding.

The best account of understanding-as-knowing-an-answer was articulated by Hasok Chang.\textsuperscript{56} According to Chang, "understanding is knowing how to perform an epistemic activity." (75) Epistemic activities he mentions include prediction (ibid.), visualization (ibid.), mechanical explanation (ibid.), counting (72), testing by overdetermination (ibid.), and narration (ibid.)—when we do these things well, we understand them. Chang's view has the major advantage that it is able to account for the fact that in many contexts mere application of an algorithm to solve a problem does not seem to count as understanding, whereas in other cases it does—it depends on what epistemic activity the would-be-understander was trying to perform. If she was engaging in the epistemic activity of prediction, she probably does understand, whereas if she was doing much of anything else she probably does not. Chang's view would predict that students who approach quantum physics with the goal of prediction would feel like they understand, whereas those who approach quantum physics wanting to visualize or mechanistically explain would feel like they do not understand—in both cases, the feeling would be accurate.

\textsuperscript{56}Chang's account would be more properly classified as an account of understanding-as-ability, since the knowledge he suggests constitutes understanding is not knowledge-that but knowledge-how. However, this distinction is notoriously difficult to draw; Jonathan Schaffer (2007) argues that all knowledge-that is reducible to knowledge how to answer particular questions, and Stanley & Williamson (2001) argue conversely that all knowledge-how is reducible to knowledge that—if either is correct, then the knowledge how that Chang claims understanders possess is not different in kind from other types of knowledge. In any event, since my reasons for rejecting his view hinge not on its appeal to understanding-as-knowledge but on its appeal to understanding-as-question-answering, it makes expositional sense to treat it here.
Despite possessing distinct advantages, there are doubts about the extensional adequacy of Chang's view. Even putting aside the controversial physics example, there are famously examples where even the best predictive know-how does not seem to constitute understanding—even if I owned the best barometer in the city and was only interested in storms so that I could predict them, it still sounds odd to say that I thereby understand storms.

A further concern with Chang’s view is that it is hard to be clear on what exactly the object of understanding is with respect to certain epistemic activities. When I explain well I understand the explanandum and when I predict well I understand the event predicted, but what do I understand when I count well? I see no connection between my ability to count the students present in my class and my understanding of those students. There might be a connection between this ability and the understanding of numbers; however, in all the other cases what was understood was object being asked about, not the method used to answer questions about that object.

In a way this last problem facing Chang is the inverse of the one facing Van Frassen—while Van Frassen (read as an account of understanding) restricted his focus on too narrow a range of questions to account for all that goes under the header of understanding, Chang's focus is so broad as to improperly count as understanding answers to the questions posed by any "epistemic activity". There seems to be no systematic way to say which questions are such that one understands just when one knows the answers to them. It is possible that there is a Goldilocks view to be had according to which understanding is knowing the answer to a contextually relevant
question of a certain type, where that type is neither circularly nor very disjunctively
defined; such a view has yet to be formulated.

The original Smithies proposal, that understanding is a particular species of
knowledge, will be a special case of the general strategy explored in chapter 6.

§2.2 Understanding as a Global Property of a System of Knowledge

On Unificationist views, explanations are not evaluated in terms of their content
but in terms of what they do to one's cognitive system as a whole. On Michael Friedman's
(1974) view, explanations reduce "the number of independent phenomena we have to
accept as ultimate." We do this by locating a few key facts from which the rest can be
logically derived. Similarly, on Philip Kitcher's (1989) view we come to better understand
the world when we can derive the most from the smallest number (yet still most
"stringent", in a loosely-defined but still basically intuitive sense) of "argument patterns",
which is a formally defined term of art.57

Friedman and Kitcher's views do not fare significantly better than the more
localized theories with respect to understanding. In at least one respect they both fare
worse—by requiring that all understanding be a function of deductive consequences, they
have ruled out accounting for non-linguistic understanding. This might be a bullet they

57Kitcher takes an argument pattern to be an ordered triple consisting of a bunch of sentences with some
non-logical vocabulary replaced by variables, a classification of which sentences are premises, which
are conclusions, and what inference rules are invoked, and a set of "instructions" for what sorts of
objects can fill in for the variables (these are essentially quantifier restrictions). How good an
explanation is is a function of how it enables us to reduce the total number of such argument patterns we
need to derive all knowledge (or observable phenomena), and of how restrictive the instructions are.
would both be willing to bite, denying that there is such a thing as non-linguistic understanding (or at least saying that, if there is such a thing, they have no interest in it).

Whatever the merits of Kitcher's and Friedman's views as theories of explanation, the most incisive criticism against them has been with regard to their relation to understanding—they are simply extensionally inadequate. Paul Humphreys, in "Does Greater Unification Equal Greater Understanding?" evokes the example of two axiomatizations of sentential logic (with modus ponens as the only inference rule), one in terms of three axioms and one in terms of one axiom. Humphreys works through the formal elements, and on both Friedman's account and Kitcher's learning the one-axiom scheme should lead to greater understanding than the three, but this does not accord with experience. Of course, experience could be fallible (see §4), but whatever the markers are for genuine increased understanding (e.g. greater usability, decreased processing time), learning the one-axiom scheme does not seem to exhibit them.

There is something highly intuitive about views that make understanding a global matter, dependent not on some local bit of information but on how it all somehow works together. That being said, cashing out precisely what this working together consists in and what end it serves has proven more difficult than one might have hoped.

---

58 1) A → (B → A)
2) [A → (B → C)] → [(A → B) → (A → C)]
3) (~B → ~A) → [(~B → A) → B]

59 <([(A → B) → (~C → ~D)] → C} → E> → [(E → A) → (D → A)]
§3 Understanding and Ability

When people understand a subject, theory, or phenomenon, they tend to possess certain abilities with respect to using or manipulating the understood thing. That much seems uncontroversial. There are two stronger claims for which one could argue, however. The strong understanding-as-ability claim would be that understanding simply is the possession of certain abilities. An intermediate understanding-as-ability claim would have it that certain abilities are crucial components of understanding, though they do not exhaust what understanding is. Both sorts of views have been explored, though, perhaps because of an inherited idea that understanding requires some degree of knowledge, the intermediate claim is more common.

Though the existence of a link between understanding and ability is prima facie plausible, this area has been explored in much greater depth only recently. Responding to a challenge from Trout regarding the value of understanding (see §4), De Regt & Dieks published an account in 2005 which included both their criterion for understanding phenomenon CUP, which turns on the crucial notion of intelligibility, and their criterion of the intelligibility of theories, CIT. Given the importance of this work in spurring the research that led to most of the accounts discussed in this section, it is worth looking at in some detail. According to the version of the account published in 2005:
CUP₁: "A phenomenon P can be understood if a theory T of P exists that is intelligible (and meets the usual logical, methodological and empirical requirements)."

(150)

CIT: "A scientific theory T (in one or more of its representations) is intelligible for scientists (in context C) if they can recognize qualitatively characteristic consequences of T without performing exact calculations." (151)

Looking at CIT immediately after CUP, one could come away from "A Contextual Approach to Scientific Understanding" feeling as if one has been given an account of what scientific understanding really is, in terms of some ability to make qualitative predictions in the relevant area of study without doing exact calculations. On closer inspection, however, one has been given no such thing. First, only sufficient conditions for understanding and intelligibility have been given. More relevantly, the sufficient condition is not on understanding at all, but only on possible understanding. All CUP₁ tells us is that something can be understood if there is an intelligible theory, but not how or by whom. Interestingly, when De Regt repeats the account in his 2009 "Understanding and Scientific Explanation", he changes CUP—acknowledging that it is different, but providing no explanation of the change. According to the 2009 version:
CUP$_2$: A phenomenon P is understood scientifically if a theory T of P exists that is intelligible (and the explanation$^{60}$ of P by T meets accepted logical and empirical requirements). (De Regt 32)

At first-glance this appears to be precisely the change we needed to get an account of (scientific) understanding,$^{61}$ but unfortunately it makes it hard to figure out precisely how CUP is supposed to combine with CIT. Combining CUP$_1$ with CIT yields the sensible thesis:

PUDD (Possible Understanding per De Regt & Dieks): A phenomenon P can be understood if there is a theory T that explains it (in accordance with accepted logical, empirical, and methodological standards) and with which some scientists in some context can recognize qualitatively characteristic consequences without performing exact calculations.

However, CUP$_2$ does not combine as well as CIT (which De Regt does not alter):

UD$_1$ (Understanding per De Regt): A phenomenon P is understood scientifically if a theory T of P exists such that scientists can recognize qualitatively characteristic

$^{60}$De Regt's reliance on explanation in his account of understanding presents no particular difficulties or circularity, as he is operating with the Deductive Nomological model of explanation (26), which is formalistic and makes no reference to understanding.

$^{61}$Along with the restriction to scientific understanding, which makes sense, and the elimination of the requirement that the explanation accord with accepted methodological standards.
consequences of T without performing exact calculations (and the explanation of P by T meets accepted logical and empirical requirements).

The problem is determining which scientists need to be able to do what. Suppose that Michael the scientist knows that T explains P, but cannot recognize qualitative characteristics of T, whereas G'Kar the scientist can recognize qualitative characteristics of T without even knowing of the existence of P (and no other scientist knows much of anything about either P or T). While this is a promising situation for future scientific exploration, it is not one wherein P is currently scientifically understood by anyone. In another 2009 piece ("The Epistemic Value of Understanding") De Regt stresses that he is only giving a sufficient condition, and that understanding is pragmatic and individual-scientist-relative, which suggests that the following is the correct reading\(^6^2\):

\[ \text{UD}_2: \text{A phenomenon } P \text{ is understood scientifically if there exists a scientist } S \text{ in context } C \text{ such that } S \text{ (in } C) \text{ can explain } P \text{ with } T \text{ and } S \text{ in } C \text{ can recognize qualitatively characteristic consequences of } T \text{ without performing exact calculations and the explanation of } P \text{ by } T \text{ meets accepted logical and empirical criteria.} \]

\(^{62}\)It does not entail that it is the correct reading, since saying that whether a scientist understands (in the active voice) is scientist-relative is more obvious and weaker than saying that whether a theory is understood (in the passive voice) should also be assessed from an individual-scientist-relative perspective. I do not think anything turns on this question.
UD₂ appears at least plausibly correct. There might be concerns about cases from quantum mechanics, wherein one possesses an explanation of an irreducibly probabilistic phenomenon in terms of Schrödinger’s Equation, which one is very adept at using, and yet one does not understand why this particular atom decayed. Still, intuitions about such cases are never uncontroversial, so perhaps De Regt (as reconstructed) has successfully given a sufficient condition on scientific understanding in one restrictive set of circumstances. We can still keep looking for a theory that applies more broadly and which, ideally, provides a necessary condition as well.

De Regt and Dieks's account, in addition to its own merits, also led to a small explosion in research on the abilities possessed by scientists who understand their field. Much of this research was conveniently compiled in one volume, Scientific Understanding: Philosophical Perspectives. In addition to UD₂, Scientific Understanding presents numerous (partial) accounts that treat understanding as an ability people have and several pieces on the connection between understanding and modeling.

In addition to Hasok Chang's view (see §2.2), another version of the understanding-as-ability thesis is presented by Petri Ylikoski in a discussion of failures of understanding. Ylikoski first says that "understanding consists of knowledge about relations of dependence" (101), but also says that "[understanding] is an ability" and that the measure of understanding is how well we can interact with the understood object (given favorable conditions). Barring some eccentric view of knowledge, the first claim appears different from the latter two. Figuring out Ylikoski's actual position is further complicated by the following passage: "When understanding is analyzed as an ability to
make counterfactual inferences in the contexts of manipulation, prediction and explanation, its relation to knowledge can be clarified. Understanding is only knowledge about the relations of dependence, not something mysterious added to the knowledge."

Barring further argument though, one might not think "only knowledge" about a particular subject matter (aside from perhaps logic itself) would never force one to make particular inferences regarding that subject matter, in which case this passage is confused—this is (part of) the lesson of "What The Tortoise Said to Achilles" (1895).

Perhaps Ylikoski is assuming some inferentialist account of meaning that makes possessing knowledge of something a sufficient condition for being able to make inferences about it.\(^63\) In any event, given that most of the examples in the paper have to do with what understanders can and cannot do, I am inclined to put aside the stated equivalence between understanding and a particular type of knowledge and treat the discussion of understanding-as-ability as Ylikoski's considered view.\(^64\)

Let us then take Ylikoski's view to be that:

\[\text{YU: One understands some object } o \text{ if and only if one can make counterfactual inferences in the contexts of manipulation, prediction, and explanation.}\]

\(^{63}\) I e-mailed this question to Professor Ylikoski. He was not concerned. Since most of his own work in this area is concerned with making explicit the practices and assumptions of working scientists, he expressed disinterest in the level of understanding possessed by stubborn tortoises who know but cannot infer.

\(^{64}\) Even if understanding-as-ability-to-make-counterfactual-inferences is not quite Ylikoski's view, it is still an interesting view that merits discussion. If his view really is that understanding actually is knowledge of metaphysically dependence relations, then it is fairly close to Woodward's causal account anyway (which account he does specifically rely on in his & Kuorikoski's "Dissecting Explanatory Power").

\(^{65}\) What, on Ylikoski's view, counts as an inference? Given Ylikoski's statement elsewhere that ""[w]hen we evaluate someone's understanding, we are not making guesses about his or her internal representations, but about the person's ability to perform according to set standards," (102) it would seem that he cannot have in mind the sort of computational model wherein inference is an operation performed on internal representations. This absence of a connection between representations and understanding marks the
This is a very plausible view. It does face a problem because of its reliance on counterfactual dependency relations as a key component to understanding—it makes it very hard to account for understanding of a necessary truth. This is particularly problematic with respect to mathematical understanding, where all (or at least almost all) of what one can understand involves necessary truths and the relations they bear to each other, and so there is no counterfactual dependence involved at all.

Finally, Margaret Morrison, while claiming "that it is neither possible nor desirable to formulate a 'theory' of understanding" (124), could be interpreted as herself endorsing a very pluralistic account. Early in her paper ("Understanding in Physics and Biology") she states that what she "mean[s] here by 'understanding' is simply having a theoretical account of how the system is constituted that enables us to solve problems, make predictions, and explain why the phenomena in question behave the way they do." (Morrison 123) One can see in this remark the start of a rapprochement between the knowledge-centric views of understanding and the ability-centric views. Morrison suggests here that understanding does involve being in a certain epistemic position, but that whether that is the right position is determined by the abilities it enables. The existence of some such complex connection between what one thinks/knows/believes/"has" when one understands and what one can do when one understands could account with much of the difficulties in producing a single, univocal sharpest difference between Ylikoski's account and my own (Chapter 5), while they share a focus on characterizing understanding in terms of its downstream effects on the understander.
account of understanding. Given this benefit, it is unfortunate that Morrison does not go into further detail. Particularly unclear on this interpretation is what counts as "a theoretical account", what counts as "the system" of which one needs to have an account, and what it means to "have" an account anyway. The question regarding what constitutes a "theoretical account" is particularly concerning. It would seem that we can understand all sorts of things without possessing a theory of them (at least consciously)—I do not think I have a theoretical account of my wife, but I am moderately confident I at least sometimes understand her. I could be wrong on either count, but I do not think the claim is conceptually confused.

§4 The Phenomenology of Understanding

Often, particularly when coming to understand something for the first time, there is a particular, pleasant sensation humans experience. Since this feeling is available for introspection, it is perhaps the most familiar and salient aspect of understanding. Nevertheless, the one fact about understanding almost unanimously agreed with by philosophers of understanding is that possession of this feeling is neither a necessary nor a sufficient condition on possession of understanding, much less in any way constitutive of it.

The distinction between the feeling of understanding and understanding itself is widely confirmed by our practice of understanding-attributions, both introspectively and with respect to others. Personally the distinction comes out most sharply when I think
back on previous reflections regarding what might be termed the deep questions of
religion, purpose, etc. For example, occasionally I would be overcome by a sense that I
had grasped some crucial religious truth (thereby coming to understand some crucial
feature of the world), only at a (usually much) later time be overcome by a sense that I
had grasped some other religious truth, incompatible with the first. Reflecting on this
incompatibility now, I am confident that on at least one of the two occasions the sense of
understanding was misleading—I had thought I had understood the world, but I had not.
Moreover, whilst in the grip of the second experience of epiphany I would deny that I had
genuine understanding in the first, so my present pattern of attributions cannot be
explained away as merely a failure to appreciate what the sensation of understanding
really feels like. Examples of this sort are sufficiently common that I do not think any
more defense needs be given to the claim that the possession of the feeling of
understanding is not sufficient for the possession of understanding.

It is equally clear from experience that possession of the feeling of understanding
is not necessary for the possession of understanding. Typically the feeling only arises, if
at all, when we come to understand some otherwise perplexing object or fact. Yet it seems
fair to say that we often understand phenomena that are incredibly familiar. To take a
trivial example, I understand objects and their permanence, even when not directly
viewed by me. My understanding of object permanence is not marked by any particular
qualitative feeling upon completion of a game of peek-a-boo. Even if one were willing to
grant that a feeling of understanding always occurred coincident with the acquisition of
understanding, that would not make it a necessary condition for or component of
possession of understanding—the possessed understanding remains long after the feeling subsides. As it happens though there does not seem to be a necessary connection between the feeling of understanding and understanding-acquisition either. Often understanding arises gradually and over the course of extended study, whereas (at least to my experience) the feeling typically accompanies only sharp discontinuous gains in understanding.

Though the feeling of understanding is neither necessary nor sufficient for possession of understanding, there is some question as to whether it is at least a reliable indicator of understanding. J.D. Trout (2002) argues that it is not, and that it is more frequently merely the expression of various cognitive biases. (Trout takes himself to be arguing that understanding itself is an unreliable indicator of truth, but, given the distinction between the feeling of understanding and understanding itself just defended, his arguments about psychological biases that can mislead the feeling are best taken to show that the feeling is an unreliable indicator of understanding.) Stephen Grimm, however, points out that most of the situations where the feeling of understanding can lead us astray are situations where we have radically mistaken background beliefs, misleading evidence, or "intellectual vices" (Grimm 2009; 90). He concludes that our sense of understanding is "conditionally reliable" (92). In "Explanation as Orgasm" Alison Gopnick argues that a non-trivial degree of reliability makes sense from an evolutionary perspective—since people who understand a situation will typically fare better in it (which better performance is probable on any account of understanding that does not merely identify understanding with the accompanying feeling), it would be
natural to evolve a more immediate reward system for understanding-acquisition. As the title indicates, the feeling of understanding evolving to provide immediate benefits for an activity that is advantageous in the long run (understanding) is analogous to the sensation of orgasm evolving to reward the advantageous activity of reproduction.

In sum, it seems reasonable to accept the possession of the feeling of understanding as some evidence of the possession of understanding, with neither its presence nor its absence being decisive. How much weight we should give that evidence is an empirical question, but one that need not be resolved before (and would likely be easier to resolve after) producing an account of what understanding is.

There is, however, one particular domain in which it has been claimed that a particular phenomenology necessarily arises with understanding—the domain of understanding one's own thoughts. David Pitt (2004, p. 26-7) argues that there is something distinctive it is like to think that p, and that the distinction between thinking that p and merely inwardly articulating a sentence expressing p ("p") is that in the former case one understands "p" whereas in the latter case one does not understand "p". Even if Pitt is correct, we cannot derive from his views any constraint on what understanding must be like—all we can conclude is that a certain mental event (thinking p) always is accompanied by understanding, and nothing about what understanding is generally accompanied by. Pitt's views do require that understanding be such that it is possible for it to accompany thinking that p, but this is next to no constraint at all—it is possible for understanding to accompany throwing a baseball or eating a sandwich. Nevertheless, Pitt's views do, if correct, suggest an added datum for an ideal theory of understanding to
explain—why is it natural to label the difference between thinking that p and inwardly articulating "p" as a difference in whether someone understands "p"?

Pitt's view on the connection between phenomenology and understanding the thought that p presents the link in its strongest form. A more moderate expression of the connection is that of Galen Strawson: "to talk of understanding-experience then, is not to commit oneself to the implausible view that there is some single qualitative type of experience that anyone who has understanding-experience must have." (2004 p. 7) To acknowledge that there is always something it is like to understand experiences (such as thinking) is not to be committed to the existence of, much less to have identified, some particular phenomenological marker of understanding.

Since on my own account understanding p consists in a particular way of representing p, when (and only when) one tokens a representation of p whilst mentally articulating "p" the possession of that representation will typically correlate with some understanding of p. Moreover, on my account of understanding language (see next chapter), the best reading of claims of the form "X understands "p" " is that X understands fact that p and X represents that fact using "p". Pitt is explicit that a key difference between thinking that p and merely internally articulating "p" is only in the former case are we using the thought representationally. Thus, both conditions of the conjunctive definition of "understands "p" " are almost always satisfied when one thinks that p, and Pitt is right to say that the difference between thinking that p and merely articulating "p" is one of understanding "p".

One might be worried that on the account of understanding just proposed there is an exception to Pitt's claim that when one thinks that p as opposed to merely internally articulating "p" one necessarily understands "p", which is when X uses "p" to represent fact that p but does not satisfy the other conditions on understanding p. On my account this is when X's representation of the fact that p is extremely non-robust (in a way discussed in the next chapter). However, it is not at all obvious that in such cases, where the content of one's thought are completely not-understood, that one is really thinking that thought at all—Pitt's own contention that we can always pick out what thought we are having on the basis of its content mitigates against this possibility. Moreover, even if it turns out that on my account of understanding we only almost-always understand "p" when using "p" to think that p, that would still explain why it seems natural to talk of the difference between thinking that p and merely internally articulating "p" in terms of a difference in understanding, which was the original datum.
§5 Understanding and Representation

One might think that understanding something seems to require more than just the possession of a particular ability—it involves a mental state underlying and enabling whole classes of abilities. One possibility is that perhaps understanding something involves representing it in a certain way. If we pick out and describe the particular way of representing that is constitutive of understanding properly, it is possible that we will end up with an account of understanding which explains understanding's theoretical and practical effects. To take an example, if understanding were constituted by the representing (by an cognizer) of the understood object in such a way that so representing it enabled one to perform the abilities typically associated with understanding, then it would come as no surprise that one who understood possessed both some theoretical information (the representation) and some ability (the ability by which that particular representation counts as constituting understanding). Objections to this view are explored in the next chapter, where an account of this sort will be defended.

§6 No Understanding?

Some have attempted to deny that "understanding" picks out a unitary feature of the world. As always, there are two ways of doing so—denying that it picks out any feature, and denying that it picks out only one feature. The former tack is taken by Jay Rosenberg,
who claims that understanding talk is best taken as expressivist—when we see people possess the ability to perform various tasks we call them "understanders", and that is all there is to the story. The latter option seems to be what motivates Morrison above, and to play a role in Paul Thagard's discussion of explanation in his *Conceptual Revolutions*. The best answer to both objections is the same: actually produce a unified account of understanding that explains our intuitions and actual use, and ideally enables progress in other areas of philosophy or even the sciences. In the next chapter I attempt to describe a unified conception of understanding that achieves these goals. However, there are a few things that can be said to ameliorate these concerns before the existence proof is produced.

While Rosenberg contends that there is no univocal notion of understanding, he takes as a starting point that no account of understanding is likely forthcoming. The role of the first half of his piece is not to demonstrate that attempts to understand understanding have failed and are likely to continue doing so, but rather to explain a pattern of failures that is simply assumed (Rosenberg 1981, p. 29). Nevertheless, we can read into his explanation of the difficulty of understanding understanding an implied argument that the task is impossible. He points out that there are three very different ways of generating understanding—compositional analysis, contextualization, and clarification. He gives examples of each sort, and one might be tempted to conclude that no mental state could properly be brought about by such diverse paths. Such a conclusion would be mistaken. First, the same type of mental state can quite often be brought about by wildly divergent means—happiness can be brought about by a memory, a perception,
a chain of reasoning, an unconscious association, and probably other things—this is not by itself sufficient reason to doubt the existence of happiness or to suppose that attributions of happiness are expressivist in nature. Second, the paths to understanding that Rosenberg presents are not nearly so disparate as they might appear—all involve acquiring new information about the understood, and in particular information that might bear on how one thinks about it.

Rosenberg's negative argument aside, it is worth taking a moment to examine his interesting positive proposal. Rosenberg attempts for understanding what has been called in other domains an inversionist strategy, wherein the customary order according to which attributions are explained by some property is inverted so as to explain the property in terms of attributions (For a discussion of this strategy, see Kraut 2007, Chapter 2). These strategies face the *prima facie* problem that, at least in the some of the domains where they have been employed, it seems possible that there could be massive and even systematic errors in attributing practice. At one time most people systematically attributed understanding of the structure of the world to alchemists, but they *were wrong* to do so; they would have been wrong to do so even if the world had ended before we could have attributed that wrongness to them. The possibility of errors of attribution implies some standard of rightness other than the practice of attribution itself. The intuition that massive error is a possibility is not sacrosanct—the inversionist can explain away this intuition without validating it. For example, the inversionist about ethics can arguably make good sense of why, given that what counts as ethical really depends on our attributive practices, we would nevertheless have good (in this case ethical) reason to
simultaneously affirm that this is not the case (for example because we want people to
treat our ethical expressions more seriously). However, this counterargument does not
automatically translate from one domain in which inversionist strategies are used to
another. Whatever our first order ethical inclinations, it makes sense that we would want
to dress up our ethical pronouncements in objective-seeming clothes if doing so would
persuade others to behave as we think they should—otherwise my neighbor would steal
my stuff, which event I would not enjoy. But what reason could I have to similarly
disguise my favored approach to epistemology? If I find particular favor in a certain
method of knowing-that or knowing-how or representing or something else, it does me no
ill turn for you to favor something else. But if there is no reason to disguise the
subjectivity of understanding, there is no alternative explanation of our intuition that
there could be massive error regarding understanding attribution.
Chapter 5: Understanding as Representation Manipulability

*Understanding* is an important notion that has been garnering increased attention, particularly from philosophers of science. In this chapter I develop several ideas from the recent literature to produce a novel account of understanding.

In §1 I examine typical judgments that employ the concept *understanding*. In §2 I review the literature on understanding from the philosophy of science, and draw out from recent discussions the importance of understanding as a state individuated by what it enables a thinker to do. In §3 I will present my account, which is (roughly) that a thinker understands when she possesses a representation of that which is understood that is sufficiently robust to be manipulable for inferential and practical purposes. I will then demonstrate the account in action, showing how it explicates characteristic *understanding* judgments, after which I will examine the role of pragmatic context in the account. In §4 I will consider applications of my account to cognitive science—specifically psychology and artificial intelligence—where I think it could be a valuable conceptual advance. Finally, in §5 I will address objections and replies.

---

67 A version of this chapter appeared as “Understanding as Representation Manipulability,” *Synthese* 190.6 (April 2013): 997-1016. Reprinted with permission.
§1 Some Preliminary Characteristic Sentences

One constraint on any theory of X is that it respect competent thinkers' characteristic judgments pertaining to X. By "respecting" such judgments, I mean (minimally) not contradicting our appraisals of the felicity of sentences expressing those judgments, and ideally illuminating the conditions under which they are true. However, a quick survey of (sentences expressing) such characteristic understanding-judgments reveals such a wide variety of claims that one might doubt that there is any single concept understanding such that an account of it could respect even most uses:

1. I understand first-order logic. (A formal theory)
2. I understand the Henkin completeness proof of first-order logic. (A proof)
3. I understand the film Chasing Amy. (An artistic object)
4. I understand the story of Chasing Amy. (A narrative--whatever narratives turn out to be)
5. I understand the assassination of Archduke Ferdinand. (A historically situated event)
6. I understand carburetors. (An artifact or mechanism)
7. I understand her. (A person)
8. I understand chess playing (The theory or practice of a game).
9. I understand Newtonian mechanics. (A physical theory)
10. I understand why the bullied teenagers reacted the way they did. (An action or actions)

11. I understand the word 'dog'. (A word)

12. I understand Ancient Hebrew. (A language)

The picture of understanding that quickly emerges is one wherein virtually anything can be the object of understanding. Most accounts of understanding treat it as a relationship between people and propositions. While that might be the most common type of understanding, taking that relationship as fundamental makes it extremely difficult to give an account of understanding, say, people or works of art. By contrast, by treating as fundamental the relationship between thinkers and objects (by which I include anything that can fall within the scope of an objectual quantifier), we can easily account for talk of understanding that $p$—typically this will just be understanding (qua thing) the fact\footnote{I do not mean to imply any ontological commitment vis-a-vis facts. States of affairs could serve the same function, or something else entirely.} that $p$.\footnote{Similarly, understanding why $p$ would typically be understanding the reason that $p$, understanding how $p$ would typically be understanding the way that $p$, etc.} Thus, this essay will build an account of what it is for a person to understand an object, with understanding propositions being a derivative case.

One might suppose that, in addition to a multiplicity of possible objects of understanding, there is an array of different relationships between thinkers and objects exemplified in the judgments above. The role of understanding in some of the judgments looks akin to a particular type of knowledge or belief—for example, a claim to
understand the assassination of Archduke Ferdinand seems remarkably similar to a claim to know a great deal about it. However, the role of understanding in other judgments looks different. Claims to understand actions could be interpreted and/or intended to indicate possession of a particular empathetic stance, and claims to understand a physical theory resemble claims to possession of a particular ability or skill.

This seeming variety of relationships between thinkers and objects might immediately suggest that it is unwise to attempt to provide a unified account of everything that goes under the name of ‘understanding’—there is simply a family resemblance, and nothing more interesting to be said about the matter. While that is always a fallback position, I contend that, surface appearance to the contrary, at one useful level of description all of these types of understanding are really of a kind—exemplars of one particular relationship that holds generally between thinkers and understood objects. Specifically, as will come out in more detail below, all are judgments about a particularly robust and interesting way the understander represents the understood object.

§2 Understanding Within and Without Philosophy of Science

To the extent that understanding has been discussed in the literature, it has typically been in the context of the philosophy of science. Philosophers of science have

---

70 There has been some work in epistemology as well, notably Jonathan Kvanvig's work on the relation between understanding and knowledge, Stephen Grimm's work on the connection between the two, and general work on the particular case of understanding one's own thoughts (see for example David Pitt's
for decades been interested in scientific explanation, and it has long been recognized that one (at least) typical result of successful explanation is the production of understanding. Historically, the approach of most philosophers of science has been to first characterize explanation and then to attempt to figure out what sort of thing understanding would have to be in order to be typically produced by successful explanations, so characterized. More recently, philosophers of science such as Michael Friedman and Philip Kitcher have developed accounts of explanation and understanding concurrently. However, it is only very recently that philosophers of science have begun taking understanding seriously in its own right as an aim of science.71

Within the philosophy of science, there are many and various views regarding what explanation is. This question is already over 60 years old in its most recent incarnation (dating from Hempel & Oppenheim's 1948 paper "The Logic of Scientific Explanation"). Current suggestions for what explanation is include nomic subsumption (Hempel 1965) conceptual (Friedman 1974) or argumentative (Kitcher 1989) unification, mechanistic decomposition (e.g. Bechtel 2008), analogical illustration (Thagard 1992 and 2002), the tracing of causal and/or dependency relationships (e.g. Salmon 1984), pragmatic question-answering (Van Frassen 1980), and a host of other things as well. All of these views face counterexamples, and all of the counterexamples are of one of two kinds: they either are examples of things the view dubs explanatory but that intuitively

71 For examples of valuable understanding arising wholly independently of explanation, see Peter Lipton’s "Understanding Without Explanation".

"The Phenomenology of Understanding"). In general though epistemology has been more concerned with knowledge than understanding. For a commentary on and explanation of this phenomenon see Linda Zagzebski's "Reclaiming Understanding". For a discussion of Grimm and Kvanvig, see next chapter.
are not, or things that the view denies are explanatory that intuitively are. But, plausibly, when we "intuit" that some bit of language is or is not explanatory, we are merely registering that it does or does not produce the sort of effect in us we would expect to be produced by genuine explanations. By fairly universal agreement, at least one major expected effect of explanation is the production of understanding; therefore, it is plausible that what one is really judging when intuining whether something is explanatory is whether or not it produces understanding. 72 If what the counterexamples really track is a dissonance between theories of explanation and intuitive judgments regarding what is conducive to understanding, then it is of paramount importance to examine what understanding is and how it relates to explanation. This importance was alluded to in the title of Michael Friedman in 1974's "Explanation and Scientific Understanding" and further defended by Jaegwon Kim in 1993's "Explanatory Knowledge and Metaphysical Dependence". However, Friedman and Kim merely assert that understanding is that which is furnished by their own favored view of explanation. As a result, they end up with accounts of understanding grounded on nothing but controversial accounts of explanation; they also fail to show how understanding could independently tie together the threads of competing accounts of explanation.

The recent interest in explicating understanding traces largely to the account of understanding found in "A Contextual Approach to Scientific Understanding" by Henk W. de Regt & Dennis Dieks. This essay argued for the conceptual importance of

72 For a more thorough defense of this claim, complete with a discussion of the most famous counterexamples, see Chapter 3.
understanding, and further suggested that understanding could be characterized largely in terms of what it enables one to do. A later collection of essays edited by de Regt, Leonelli, and Eigner explored the form of understanding as attributed in various specific sciences.

My own account, formulated in the next section, has two central claims: first, understanding is at root the possession of the right sort of mental representations of that which is understood; second, that a mental representation counts as being "of the right sort" in virtue of the fact that possession of it enables one to perform (typically, but not always, intellectual) feats relevant in that context. In the remainder of this section I will discuss two important precursors to this account found in the de Regt et al anthology, but also argue why these precursors are not well-suited to be expanded to an account of understanding generally.

The importance of the actions and/or inferences putatively enabled by understanding is stressed by de Regt. Specifically, he argues in defense of the following claim:

In some context C, the existence of an explanation of the occurrence of a phenomenon from some theory T such that scientists in C can "recognize qualitatively characteristic consequences of T without performing exact calculations" suffices for that phenomenon being understood.73 (de Regt follows

---

73 An earlier version of de Regt's account from de Regt and Dieks's "A Contextual Approach to Scientific Understanding" is weaker, providing only a sufficient condition for when a phenomenon "can be
Hempel in regarding explanation as an argument from a theory to a statement of a particular phenomenon (de Regt 2009, 32))

While this account is extremely plausible in certain contexts, it would have several limitations if taken as the basis of a general account of understanding. First, the account is very domain restricted—it only gives a sufficient condition on one type of understanding in scientific contexts where calculation is especially relevant. Second, de Regt's account is surprisingly impersonal—it gives a sufficient condition for phenomena being understood, but no general account of what it is in virtue of which a subject understands. In a situation in which some scientists can recognize characteristic consequences of a theory and others can explain a phenomenon using that theory it is not clear who, if anyone, understands what. Third, the account demands a fundamental asymmetry between understanding phenomena and understanding theories—the two types of understanding are discontinuous, and one can only understand the former via understanding the latter. But while it is certainly common to come to understand a phenomenon in virtue of understanding a theory that predicts it, it also seems that sometimes we come to understand a theory in virtue of understanding the phenomena it predicts. The ability to represent Newton's laws of motions enables me to understand this rock in free fall. However, understanding this falling rock can also help me understand

understood" (de Regt & Dieks 2005, 151). For a discussion of the implications of this change, see Chapter 4.

For more on this point, see Chapter 4.

Theories are understood in virtue of being intelligible, whereas phenomena are understood in virtue of being explained by theories.
Newton's theory, which fact is utilized in classroom demonstrations of such laws. Fourth, it would be a theoretical boon if we could formulate an account of what it is to understand without relying on a notion of explanation, as what explanation is is the subject of much dispute in the philosophy of science, the resolution of which dispute might hinge on what understanding is.\textsuperscript{76}

Another difference between de Regt's account and my own (developed below) is the latter's emphasis on the understander possessing a representation of that which is understood. The idea that one's understanding x consists in large part of representing x in the right sort of way is not wholly new either. However, where I will contend that "the right sort of way" is best cashed out as a mental representation the possession of which enables certain abilities, it has been more customary to suggest that understanding is the product of the possession of a model.\textsuperscript{77} Models are (often physical) objects that bear enough structural similarities to the thing modeled to enable surrogative reasoning—that is, are such that one can justifiably infer from one's conclusions reached while reasoning about the model to some fact about the thing modeled. I claim that it is preferable to build an account of understanding out of mental representations rather than models for two reasons. The first reason is that when I understand cellular meiosis, what I am reasoning about is precisely cellular meiosis itself rather than some structurally similar other.

\textsuperscript{76}Michael Scrivens first put forward the suggestion that understanding be taken as a constraint on explanation in his "Explanation, Prediction, and Laws". In Chapter 3 I argue that not only can a general account of understanding be constructed independently of an account of explanation, but that the former can actually be an invaluable tool in our conceptual repertoire when attempting to explicate the latter. The philosophy of science thus provides both the most sophisticated theories of understanding in the philosophical literature, and the most obvious potential consumer of a new theory.

\textsuperscript{77}See e.g. Knuttila & Merz (2009) or Lenhard (2009).
Representations, by their very nature, capture this intentionality—what it means to be reasoning about meiosis is just to be engaged in operations on the appropriate mental representations. When one talks of models, however, not only is this representational aspect de-emphasized, but its relevance and even existence is sometimes explicitly denied (see §5). The second reason to build the account out of mental representations instead of models is that understanding is not just constituted by my thinking about cellular meiosis, but also by my thinking about cellular meiosis. This distinction is most clearly drawn by contrasting possession of a mental representation with possession and use of a merely physical model. I could blithely follow an algorithm from a manual to operate a model airplane without thereby understanding planes. This wedge that can be driven between possessing a model and understanding cannot be as easily driven between possessing a mental representation and understanding.\footnote{One could produce an account of understanding as the possession of specifically mental and representational models, in which case the difference between the talk of such models and mental representations would be negligible.}

§3 The Positive Proposal

In this section I put forth my explication of understanding. The account is built up and then demonstrated in action in §3.1. In §3.2 I look at the role of context sensitivity.
§3.1 The Account

Following de Regt and Dieks, I take the critical component of understanding to be its empowering role—the difference between understanders and non-understanders is that the former, but not the latter, can utilize the understood effectively. But of course various factors prevent or empower one to affect things in the world without being signs of understanding or its absence. Inability to do derivations in first-order logic could arise from a broken pencil; conversely, ability to perform such a derivation could result from having memorized this particular derivation or just being extremely lucky. A wide array of factors that are not cognitive, much less markers of understanding, can affect our ability to interact with the world. Understanding is a cognitive achievement, and so no non-cognitive ability is likely a conceptually necessary result—instead of looking for what understanding empowers one to do in the world, then, we would do well to look for a cognitive effect that prefigures worldly abilities. This suggests that understanding must instead consist of an ability to manipulate some mental correlate of the understood object such that, in the absence of interfering factors, one would then be able to manipulate the target itself—the obvious mental correlates of objects for a given understander are that understander's mental representations of those very objects.

The essence of the account is that one understands x when one can manipulate a representation of it in the right sorts of ways. Before spelling out what those are, it is best

---

79 This is similar to James Woodward's manipulationist account of explanation, according to which one explains by citing causal variables the alteration of which (in appropriately specified counterfactuals) would have affected the explanandum.
to be explicit about a substantive, albeit relatively uncontroversial cognitive scientific assumption:

(A) In order to understand some object x, a thinker must possess a mental representation of x.

I would like to stay as neutral as possible regarding what representations are. Minimally, they are computational structures with content that are susceptible to mental transformations. That we have mental particulars that satisfy this characterization is consistent with classic computationalism, a language of thought hypothesis, and at least some (non-eliminative) interpretations of neural networks. Such possession is not consistent with eliminativism about the mental. Since a majority practitioners and philosophers of cognitive science are committed to the existence of some sort of explanatory efficacious representations anyway, the project here is an attempt to show that that state to which most are already committed can pay explanatory dividends in the realm of understanding as well.

Combining the intuition that understanding an object gives us certain abilities to engage in manipulation with the assumption that for every understood object there is a corresponding representation, the account of understanding takes shape:
URM (Understanding as Representation Manipulability): A statement, attributed in context C, that thinker T understands object o, is true if and only if T possesses a mental representation R of o that T could (in counterfactuals salient in C) modify in small ways to produce R', where R' is a representation of o and possession of R' enables efficacious (according to standards relevant in C) inferences pertaining to, or manipulations, of o.

By requiring that understanding consist in a particular way of representing the understood object, URM accounts for both the general intuition that there is a specifically cognitive aspect of understanding and the specific intuition that someone who merely "gets lucky" and behaves in all the ways one would expect an understander to behave does not thereby necessarily understand. However, since, for each particular representation, whether possession of it will count as understanding will depend on what one can do with it, we can also account for both the contextual variability of understanding and its predictive uses. By further requiring that one be able to modify the representation itself, URM rightly categorizes mere application of a memorized algorithm as not genuine understanding.

80 For simplicity, I assume that all understanding statements are uttered and evaluated in the same context—what I call the "context of attribution". The precise semantics and pragmatics of understanding claims are well beyond the scope of this essay.

81 While I do assume in §5 that there must be some objective basis for R counting as a representation of that o (any of similarity, isomorphism, or causal connection will due for those purposes), it certainly need not be a perfectly accurate representation.
Working through the way this account explicates some of the sample characteristic judgments from §1 will both make it more concrete and provide evidence that URM sheds genuine light on what makes such judgments true or false.

Starting with #1, my understanding first-order logic consists in my possessing a representation of first-order logic with which I can do certain things. I represent logical theory by a collection of representations of semantic properties and of allowable rules in natural deduction. I could notice and correct inconsistencies or mistakes in my representation; had I somehow learned the incorrect rule for &-elimination, my appreciation of the semantics or of the harmony between introduction and elimination rules would enable me to spot the mistake and fix it (thus creating a new representation R').

Understanding the Henkin proof of the completeness of a system of natural deduction of first-order logic works similarly: I possess a representation of the proof that can be manipulated to either correct small errors or to produce a new proof of some

---

82 What constitutes a representation of a proof will depend largely on how the proof is conceived of by a given understander. Thus the understanding of a mathematician who thinks proofs are abstract objects might be very different from the representation of someone who conceives of proofs as particular sequences of sentences. These two conceptions will result in very different uses of the concept proof and very different representations. If there is a fact of the matter as to what proofs really are, then the content of both representations will need to be sufficiently similar to that fact in order for both to count as representations of proofs, though one representation might still be more accurate. We would typically expect the more accurate representation to be more useful and hence, per URM, the person with the more accurate representation to understand better (all else being equal). This seems like the intuitively correct result. However, there might be occasions where certain idealizations and/or abstractions make a representation less accurate but more useful, in which case the person with the less accurate representation would be expected (both intuitively and according to URM) to understand more, so long as her representation was sufficiently accurate to still underlie understanding of that specific proof. (No matter how useful a representation of pajamas might accidentally be, it still could not be the basis of understanding the Henkin Completeness Proof).

83 Since the proof is valid, any errors to be corrected would be in the representation of the proof, not the proof itself. Nevertheless, if the proof were not valid and one could see how to fix it, then that would be even more impressive understanding (relative to most contexts). Whether the improved understanding
similar theorem (such as the completeness of a slightly different natural deduction system, or a system for some different semantics)\textsuperscript{84}. I once learned in a class, but very much did not understand, a proof of the completeness of Frege's extraordinarily non-natural deduction system. I had a representation of that proof too—I could conceivably have replicated it on an exam. However, that representation was not manipulable by me—I could not possibly have used it to prove the completeness of an axiomatic deductive system with even slightly different axioms, much less a wholly different sort of logic.

Understanding a specific event, such as the assassination of Archduke Ferdinand, is just like understanding an object, and so poses no particularly new problems. This lack of a distinction between human behavior and scientific facts is only notable because of the presence of such a distinction between \textit{Verstehen} and \textit{Erklären} (generally translated as \textit{understanding} and \textit{explanation}, respectively) found in the Continental tradition. On the present account, there is no need for such a sharp dichotomy.

Since carburetors are mechanisms, one would expect understanding of how carburetors work to look a lot like mechanistic explanation. Typically, it will; one will be able to figure out how to fix or improve this carburetor if one's representation of it includes a decomposition into functional parts and contextualization regarding its role in

\footnotesize
\begin{quote}
\textsuperscript{84}Fully accounting for the case of applying a variant of the proof to a new object requires making sense of some "core" of the proof that was understood in both cases, such that both my original representation R and my new representation R' of the new system are still representations of that. This requirement is not burdensome, for if there were no such shared core at all, be it even stylistic, there would be no commonality in virtue of which we could call the different proofs variants of each other.
\end{quote}
some broader system.\textsuperscript{85} This tight fit between the prediction of URM (in standard circumstances) and a plausible account of explanation and understanding within a particular narrow domain gives more reason to suppose that URM is a plausible account of understanding more generally.

"I understand her" in virtue of representing her in a particular way, where that representation is robust enough to be manipulable and provide good inferences about her past or future actions. One could also intend by this same sentence to express the judgment that one empathizes with her, in which case this judgment is analogous to the one regarding the troubled teenagers discussed below.

If "I understand Newtonian mechanics" is read as a statement about a particular theory, then it can be treated similarly to a statement about understanding first-order logic. However, if the sentence is interpreted to indicate the possession of some mechanical aptitude, it is still readily made sense of in light of URM; it is a strength of URM that it shows just how the two readings are related. Where the first reading posits a manipulable representation of a particular theory, the second posits a manipulable representation of a particular class of objects and their relations. The know-how required to, say, launch a cannonball, involves at least in part the manipulation of a representation of that cannon firing. I picture where the ball would go if launched from various orientations, determine which modification to the representation would result in optimal launch conditions, and then make the corresponding alteration on the actual cannon. Assuming such manipulation does occur, it shows us what is the same in the seemingly

\textsuperscript{85}See Bechtel's 2008 \textit{Mental Mechanisms} for an account like this.
different achievements of understanding some particular linguistic theory and understanding the non-linguistic phenomenon that theory was intended to model.

The case of understanding the reaction of the bullied teenagers is curious. If it is simply the expression of an epistemic relation, it says something about how one can manipulate one's model of the teenagers in their environment in such a way that the behavior in question would naturally follow. The empathetic reading goes a bit further—one could by this sentence express empathy with their actions. At first this might appear to be a straightforward ambiguity in our term 'understanding', for empathy does not seem cognitive in the same way as most uses of 'understanding' do. But empathy requires as a prerequisite a certain cognitive achievement, the ability to "put yourself in someone else's shoes"--and that achievement is exactly the sort of thing URM would say constitutes understanding. What one does when empathizing is modify a particular representation of that with which one is empathizing into a representation of one's self (or at least of someone one finds morally or otherwise sympathetic). Thus, part of empathizing with the boys is picturing one's self being bullied in the same way, and realizing that one might have reacted similarly. Thus at worst the empathetic reading of this sentence indicates a minor polysemy of 'understanding'. Again, it is a strength of URM that it shows how two seemingly very different interpretations of a sentence attributing understanding are related.

Understanding a language is a complicated case. There are two routes open here. In some contexts it makes sense to say that understanding languages and linguistic items is just like understanding any other objects, in which case understanding 'p' would
involve being able to manipulate the object (or a representation of) 'p'--with some filling out, this approximates a meaning-as-use theory of understanding language. However, in other contexts understanding language seems to consist of something slightly different—in such contexts what someone means by "John understands 'p'" is just that John understands that p (and represents that p using ‘p’). Before questions of understanding language can be approached, however, it is necessary to say a bit more about what it means for a word to have the semantic value it does—of particular importance is the question of whether a competent language user possesses an explicit representation of the links between language and world, which would suggest that the first route proposed (i.e. that we possess manipulable representations of language itself) might apply fairly generally. Producing a complete account of understanding language is a complex endeavor that merits full treatment elsewhere.

In addition to the foregoing, we can also use URM to make sense of understanding URM itself. I have a representation of URM which can be modified or adapted to all sorts of different circumstances and sentences regarding understanding, as evidenced in the preceding paragraphs. This compares favorably to other accounts, many

---

86 For an account with this feature, which makes URM a bit easier to apply, see for example Larson & Siegal's Knowledge of Meaning.

87 As with representations of mathematical proofs (see n. 15), individuals' representations of a philosophical theory likely vary with their broader views regarding what philosophical theories are. For example, a realist might suppose that philosophical theories give information about underlying facts. In that case, a representation of a theory like URM would be like the representation of a baseball box score—a complex representation comprised of representations of some bits of language and of the semantic facts those bits provide information about. I stress that this is just an example of what one possible structure of a representation of URM, in order to show how application of URM can be made concrete—an evaluation of the plausibility of such a picture is well beyond the scope of this essay.
of which do not apply to themselves.\textsuperscript{88} This inapplicability need not be grounds for criticism, as there is no particular requirement for understanding to be within its own domain, and many accounts explicitly restrict their domains of interest in some way or other. Still, it seems like we should be able to understand understanding, and that any account that can apply to understanding itself is richer on that account.

§3.2 Context-Sensitivity

The case where I can correct mistakes in my representation of first-order logic and apply first-order logic to lots of new and interesting problems is fairly clearly a case of understanding classified as such by URM. However, what should be said about more questionable cases? For example, does someone who can apply the rules of first-order logic, but who cannot do any meta-theory or even articulate clearly what she is doing count as understanding first-order logic or not? There is not obviously a correct answer to this question. Focusing on the target phenomenon—our judgments and attributions

\textsuperscript{88}For example, on Michael Friedman's account, we understand when we can reduce the total number of independent facts we need to take as primitive. If we had a number of viable theories of understanding, each requiring its own independent facts and each of which was subsumed by Friedman's account, then his account might help us understand understanding. But as that does not seem to be the state in which we find ourselves, it does not do so.

Other accounts are so geared towards understanding in science that it is not clear how one would even begin to self-apply them. On Carl Hempel's D-N account, for example, one understands something when one could have deduced from one or more laws of nature (and zero or more initial and boundary conditions) that it had to be true. One could not treat the truth of his account as an event whose occurrence (when did it happen?) could have been predicted—the alternative is that his account is itself a law derivable from a more general law. I do not know what the domain of such a more general law would be. It could not be a law governing mental states, since understanding per Hempel is non-psycologistic, and there is no where else obvious to look. (For further details regarding other accounts, see my chapter 2).
pertaining to *understanding*—there is quite a bit of context-sensitivity. The understanding of someone who got a 4 on her AP calculus test might clearly count as understanding when being evaluated for one job opportunity (e.g. high-school summer intern), clearly not count when being evaluated for another (e.g. professorship in an elite math department), and neither clearly count nor not count when being evaluated for yet a third (e.g. admission as a student to an elite math department). The target notion thus exhibits sensitivity to context.

How can different contexts affect the appropriate application of our common-sense notion of understanding? There are at least two dimensions of variability—which sorts of attributes matter to determining whether something counts as understanding will vary contextually, as will the degree to which those attributes must be present to be above some threshold of understanding.

What sorts of attributes, in what contexts, count as relevant to determining degree of understanding? Intuitively, the sorts of attributes relevant to determining degree of understanding in some particular context are those that enable one to make the types of inferences and perform the types of manipulation that are relevant in that context. In a class on meta-logic meta-theoretical inferences will be relevant, and someone who can make more meta-logical inferences and produce more meta-logical proofs will count as understanding better even if she is not particularly skilled at object-language derivations. In an introductory course on proof construction, the ability to construct object-language proofs will be relevant and a student who can perform more kinds of object-language
proofs will count as understanding better, even if she cannot say anything about the meta-
theory behind her behavior.\(^89\)

There is at least one other dimension to our common-sense notion of understanding: the extent to which one understands varies not just with what one can do, but the conditions under which one can do it. Suppose Vir is a brilliant psychologist but markedly socially inept, whereas Londo is a social butterfly who could not put two sentences together to explain how people behave (much less why). When seeking a speaker for a conference on the psychology of party-goers, one would clearly want Vir; in a context where we are looking at applications to speak from Vir and Londo, it would be fair to judge that Vir understands people, whereas Londo does not. However, matters differ sharply when sending out invitations to a party. We are well aware that Vir will be awkward and never know how to interpret people or how to respond appropriately. In that context, it seems reasonable to judge that Vir does not understand people, whereas Londo does. Differing contexts thus exhibit discontinuous criteria regarding what counts as understanding, even of the same object. We could try to locate the context sensitivity entirely in the specification of an object of understanding, but this case demonstrates why that would be difficult—the object of Vir and Londo's respective understandings can be arbitrarily similar (people-at-parties-when-approaching-the-keg-from-the-left...). Unless

---

\(^{89}\)The relevant context, moreover, appears to be the context of attribution. While it seems reasonable to suppose that an instructor of a proof construction course might truthfully say that Clark understands logic even as an instructor of a meta-logic course (in a different conversation) truthfully denies this, it would be jarring if, for a single speaker, the claim "Clark understands logic" could change truth values as Clark moves from his 2:30 proof construction class to his 3:30 meta-logic class. There are many subtleties left untouched regarding where in the linguistic machinery context has effect, but, as already mentioned, a detailed semantics of **understanding** is well beyond the scope of this essay.
one builds a particular perspective into the specification of an object, the same object—
described to arbitrarily fine grain—can be viewed from different perspectives. Hence it
will be possible for the very same object to be understood in some contexts but not in
others. By the same token, the example does not rely on any particular ambiguity in
sentences of the type "X understands people"—one could see the same sort of contextual
variance (for exactly the same reason) regarding the question of whether Vir understands
why the beer at parties always seems fresher in the second keg tapped than the first.

In order to capture the target phenomenon, an explication of understanding should
be expected to incorporate context-sensitive parameters in such a way that they line up
with their counterparts in the target. URM does so.

To begin with, I will discuss how URM captures contextually variable thresholds
of understanding. For example, a student who could perform new logic derivations by
applying the known rules to very slightly different examples might count as
understanding, or might not. That is the prediction of URM: taking o as first-order logic
generally, R as some particular representation of schematic rules, and R' as a version of
the rules that applies to the most difficult logic problem that student could solve, one who
wants to determine whether the student understands o must first precisify what counts as
an "efficacious inference" about o. The ability to make enough inferences to get a C+ on
a well-designed exam is (some) evidence for possessing more understanding than the
person who can only get an F but less understanding than the person who easily gets an
A. Whether the particular inferences are sufficiently efficacious to count as understanding
* simpliciter is a question without a determinate answer. Context-sensitivity lurks here as
well: what counts as efficacious might differ for a first year student, an honors student, and an applicant for a tenured position.

Context-sensitivity enters into the machinery of URM most explicitly and most crucially in the counterfactual nature of URM: what is the hardest problem the student could solve? Returning to Vir and Londo, the restriction to salient counterfactuals explains the difference in their understanding⁹⁰. When Vir is surrounded by loud and bustling party-goers, he is simply not capable of updating his representations well enough or fast enough to be able to interact effectively with the world. Of course in the counterfactual scenario where Vir is sitting in a comfortable chair at a conference discussion panel, he would be able to represent the theoretical party-goers very well and make all sorts of interesting inferences regarding their behavior; however, whatever salience amounts to, that counterfactual is not typically salient during an average party. We could imagine cases where it becomes salient, but those are the very same cases where the correctness of judgments regarding his understanding seem to shift—were someone to ask specifically about Vir's academic credentials, one could very well imagine being told, even as Vir commits yet another faux pas, that Vir really understands people. This example shows that URM also lines up with the target notion in locating

---

⁹⁰I would like to remain neutral regarding how precisely the restriction to salient counterfactuals informs the semantics of ‘understanding’. The most straightforward way would be to take contextually salient factors as defining a set of possible worlds according to some function. The understanding relation itself would then be an ordered-triple featuring a person, an object, and a set of possible worlds (the elements of which are themselves picked out by a function from contexts to worlds). Thus, for example, <Vir, people, {x: x is a world at which Vir is speaking at an academic conference}> is in the extension of ‘understands’, and the sentence "Vir understands people" is true in contexts in which the restriction to those worlds where he's speaking at an academic context is salient. There are doubtless other ways to incorporate the context-sensitivity.
variability in the context of attribution, rather than in the context where Vir happens to find himself (see n. 89). We can suppose that Londo, by contrast, lacks the ability to represent people's psychology while sitting at a conference table—he needs the immediate contact to spur his thoughts.

§4 URM in the Field

§4.1 Understanding in Psychology

In the psychology literature, there is interesting research regarding the downstream cognitive effects of explanation possession. For example, the works of Barbara Koslowski and Tania Lombrozo detail the cognitive impact of hearing or possessing or coming up with an explanation. However, there is almost no discussion of understanding in this literature, and I contend that much of the reason for that is that understanding has until now been too ill-defined a concept to be helpful in psychological research.

Here is an example where a notion of understanding like URM might have been helpful to psychology. In Prof. Koslowski's experiments, students were presented with an explanation of something and then tested to see whether possession of that explanation changed what sorts of facts counted as evidence relevant to the domain in question. It did—in two out of three experimental conditions. In one of those conditions, an explanation was presented that detailed a more precise (relative to the other two
experimental conditions) mechanism that produced the fact under investigation. In the other successful condition, they were reminded of the explanation at just the right moment. However, in the third condition, where they neither possessed knowledge of a detailed mechanism nor were primed to remember the explanation, having been told of some explanation had no significant effect. What this suggests is that there is a causally salient step between merely receiving an explanation on the one hand and being able to employ it properly on the other. The priming condition might suggest that all that is required to get the appropriate results is to remember the explanation at the right time, but the detailed-mechanism condition suggests that this priming only supplements something (which we could just call 'understanding') that could have been caused by the right type of explanation in the first place. Perhaps what her experiments demonstrated was that there was a crucial step—understanding—mediating between the explanations given and the results of the studies. When the potential utility of understanding was suggested to her, Koslowski responded by pointing out that there was no clear understanding of understanding that could be utilized in this case. That is something that could conceivably be remedied by URM.

I am hesitant to say exactly how a skilled experimenter could make use of URM. However, if I am correct that the difference between Koslowski’s mechanism and non-mechanism conditions is that only in the former did subjects tend to gain understanding, and that URM is the correct account of understanding, then one would expect subjects in the former but not the latter condition to possess particularly robust representations of the

---

91 Personal correspondence.
situation described. This possession of such representations seems in principle testable; two routes for doing so suggest themselves. First, one could ask the subjects other questions about the stories, not pertaining to what they would count as evidence, the ability to answer which would typically be associated with subjects having filled-out "pictures" of the stories. Second, if we could identify particular brain areas, the activation of which was proportional to some modality of representing (e.g. visualization), and that area showed significantly more activation in the mechanism condition, then that would be some evidence for my hypothesis as well.

Another approach to understanding in the psychology literature is to treat concept-acquisition and application as paradigmatic or even constitutive of understanding the world (so conceptualized). Psychologists typically consider concepts to be mental particulars that are representational and according to which we categorize objects in the world. URM predicts that concept acquisition (where concepts are understood as just specified) would, in general, be a means of gaining understanding. So long as one is in a context where categorization, or some other use of concepts, would enable efficacious inference, concept acquisition entails that one acquires a new, useful representation—exactly the sort of process that would constitute understanding per URM. Thus psychologists' reliance on concept acquisition as understanding could be seen as merely one particular instance of URM.
§4.2 Understanding in Artificial Intelligence

Artificial Intelligence is perhaps the field most in need of a clean conception of understanding, and the field best positioned to make use of URM in particular. Aside from its intrinsic interest as a cognitive state, if one wants to build a computer that behaves like a human cognizer, it helps to be as clear as possible about what exactly human cognizers are doing when they claim to understand something. URM reduces understanding—a notion that picks out a particularly complex cognitive state—to representation, inference, and object manipulation. While these latter three concepts are also complex, they are far more familiar to cognitive scientists. Cognitive science is extremely occupied with the study of representation, and programs for modeling inference and object manipulation have existed for decades.

Context and counterfactual salience might be a bit less transparent to modern cognitive science than representation, inference, and object manipulation, but it is important to remember that the counterfactual and contextual elements of URM appear only in the characterization of what it is for someone to understand—outside the scope of what the understander him/her/itself needs to represent at all. Vir understands in virtue of his ability to manipulate certain representations in order to make certain inferences or manipulations; which such manipulations will constitute understanding is contextual and

92 According to Bobrow & Brown "Good representation is the key to AI", as cited (admittedly with disapproval) by Brooks (1991, p. 141).
93 Such as Newell & Simon's GPS (2000, p. 84).
94 Paul Churchland has a useful presentation of the work of Llinas and Pellionisz in Chapter 5 of his A Neurocomputational Perspective.
depends on the truth of certain counterfactuals, but Vir himself need not know, believe, or otherwise represent that fact. He need only be able to modify the representations and make the inferences or perform the manipulations.

§5 Objections and Replies

Objection 1: Understanding Ain't in the Head!

Objection (version A): Petri Ylikoski, responding to a suggestion by Jonathan Waskan to the effect that we should understand understanding in terms of modeling, says that this "cannot be the whole story about understanding. When we evaluate someone's understanding, we are not making guesses about his or her internal representations, but about the person's ability to perform according to set standards" (2009, p. 102).

This Wittgensteinian objection, if applied to URM, seems ill-founded for three reasons. First, it does not follow without further argument that one can read off from our practice regarding attributions of understanding that understanding is not an internal affair. If understanding were a wholly internal affair, one might still anticipate that attributions of understanding would depend on external criteria.

Second, while Ylikoski asserts that we are not making guesses about people's representations, this claim is not argued for. I have had students claim that while they understood the material they were still for some reason unable to perform certain basic tasks with the theory. While I am frequently skeptical of the truth of such claims, neither
the claims nor the students seem to be conceptually confused. Even when I reject the students' claims, I am still hazarding a guess about what is going on in their minds. And in some cases I actually do believe that they understand. Consider a dyslexic student. Universities (at least in the United States) make special allowances for students with such conditions as dyslexia precisely because it is recognized that they cannot always perform the tasks we come to expect of people who understand the material. One might be cautiously optimistic that there are other tasks one could ask them to perform such that their performance would be reflective of genuine understanding, but it would be foolhardy to claim that this is necessarily or even always actually the case.

Third, if one were to try to apply this objection to URM, the claim that what we are guessing at is merely a representation in someone's head would be misleading. According to URM we are attributing to putative understanders a particular kind of representation, but one that is individuated by what it enables one to do. Thus, in just saying that they have a representation of this particular type one is also saying that there is something they can do (albeit possibly just something mental).

Objection (version B): But URM is entirely internalist! All the criteria depend on what is merely happening inside one's head, and that cannot be right. Genuine understanding requires some accurate fit with the real world.95

95JD Trout, for example, claims that understanding is a matter of "being in the right cognitive relation to the evidence," (2004, p. 202) and that the "sense of understanding" is merely a reflection of intellectual biases (2002). If URM identified understanding solely by features as perceived by the understander, then it would risk being subject to such biases and disconnected from the real world. However, while URM does not take understanding to be a fundamentally evidential relationship, I argue that it does not rely on wholly internalist criteria either.
**Reply B:** URM does not rely solely internalist criteria; the extra-mental world enters the account in two ways. The first constraint on understanding imposed by the world is the requirement that R and R' actually represent objects. Whatever representation amounts to, there is probably some worldly component such as, at a minimum, reliable co-variation between that which represents and that which is represented. The second non-internalist constraint is the requirement that the manipulations enabled by generation of R' be efficacious. If R' enables new inferences about o', then it is required that these inferences be good ones, which will at least sometimes require reference to the world for background conditions. Obviously if the benefit derived from some particular R' is some new ability to manipulate o, then there will be at least some worldly check on understanding there as well (it is difficult to successfully manipulate an object if one is wrong about it being, say, smaller than a galaxy).

**Objection 2: Nonrepresentational Models**

*Objection:* Knuuttila & Merz (2009, p. 148-9) pose the following dilemma for the theorist who considers models representational: either representations represent because of some objective similarity to the represented object or they do not. If they do, then, as similarity is a symmetric relation, the represented object also represents what we had been taking as the representation, which is absurd. If they do not, then their status as representations is determined by the users' intentions. But if representations are representational merely because they are intended to be so and not because of any intrinsic characteristic, then we
could not possibly examine representations to learn about the represented phenomenon. If
the relation between the representation and represented object is arbitrarily fixed by the
user (the "pragmatic" conception attributed to philosophers such as Suarez), then there is
no way possession or examination of such a representation could teach us anything about
the represented object, and so it is not clear how such possession or examination could be
any part of understanding. Either way, representing something cannot be the key to
(much less constitutive of) understanding it.

Reply: There is a way out of this dilemma. One can grant that a representation's status as
representational is determined by users' intentions without thereby granting that there is
never any useful degree of similarity between the representation and the represented. We
sometimes select certain representations precisely because of some relevant dimension of
similarity. I could use a salt shaker to represent Lisa del Gicondo, but if my goal were to
study her appearance it would be much wiser to represent her with a painting. The
painting's status as representation is determined pragmatically, but it does not follow from
this fact that the painting does not look like Lisa. The reason selection of models and
other representations is so hard in science is because it can be done either badly or well
from the perspective of choosing representations that will be useful for study.

Objection 3: Too Loose

Objection: There are so many vague and context-sensitive terms in URM that a
sufficiently motivated theorist could make virtually any understanding-claim come out
true. The exercise of application in §3.1 is less proof of the correctness of URM than it is the of URM's theoretical promiscuity.

*Reply:* This is a very real concern, which can best be answered by demonstrating that various understanding claims do not come out true according to URM. It should however be granted at the outset that URM is not as precise as (for example) a mathematical definition, and so it will be difficult or impossible to *prove* that specific understanding claims can never be true in any context. It is however fair to demand that some understanding be classified as false in typical circumstances (in a non ad-hoc way).

Let us start with the easiest case. If I had never seen or heard of linear algebra, I clearly could not have understood it according to URM—there would have been no representation R to manipulate. Hence, a claim that I understood linear algebra in such a context would come out indisputably false. Next, if I had heard the name "linear algebra" and even knew that it was a something mathematics-y (without possessing a substantive theory of what this amounts to), I could not engage in any manipulations. I could not apply any representation of linear algebra that did not represent it as being mathematical to anything effectively, since the only aspect of linear algebra I represent is its being broadly mathematical. Cases like these show that URM is not vacuous or trivial.

Problems arise, however, when we add complexity to my representation of linear algebra. As soon as a case is even moderately complex, it likely becomes possible for a sufficiently creative philosopher to come up with contexts where that bit of representational flexibility would be just what one needs for a genuinely productive
manipulation. At this point there is little alternative but to just try to apply URM in specific contexts and hope that the contextual sensitivity in URM lines up with our intuitions about the proper application of the term 'understanding' in each context. URM is at least contentful enough to have implications for evaluating extant views in the literature on scientific explanation, on the assumption that explanations should at least often result in understanding. For example, Deductive-Nomological (D-N) explanation and the sort of understanding Hempel claimed it could provide (Hempel 1965, 337) do not seem to fit a URM framework at all. John could very easily manipulate his representation of an event without having the slightest notion what law of nature subsumes it, and it is at least plausible that the ability to perform certain derivations will not typically give him any new manipulating-ability with respect to representations of the derived phenomenon. Thus if URM is the right account of understanding and the link between explanation and understanding is granted, D-N is likely not the correct view of explanation. (It is true that knowledge of laws governing an event would usually be at least some help in enabling one to manipulate a representation of it, which fact might explain part of D-N's appeal.)

§6 Conclusion

Accepting URM as the correct account of understanding would shed light on the truth conditions of our understanding claims, and potentially be a valuable advance in cognitive science. There are still details to be worked out, particularly regarding the
pragmatics of understanding claims, the details of what constitutes salience, and perhaps some criteria for individuating representations, but, overall, URM is a decided step toward a better account of understanding.
Chapter 6: Understanding sans Justification

In this chapter, I argue that considerations from the philosophy of explanation and understanding yield surprising epistemological fruit. The correctness of URM will be sufficient to generate the results of interest, though I will attempt to keep the present chapter as neutral as possible regarding the details of the correct account of understanding. The chapter centers around the inconsistency of the following three claims:

1) Not all instances of understanding hold in virtue of possessing justification.

2) Understanding is a species of knowledge.

3) Any knowledge one possesses obtains at least in part in virtue of one possessing justified beliefs.

Here is the structure of what follows. In section 1 I explore the relationship between understanding and knowledge, and restrict the focus of the discussion to the narrower question of understanding states-of-affairs. I also provide a reason to think that the question of what understanding is or in virtue of what understanding obtains are of general philosophical interest. The original contribution of this paper is predominantly in §2, where I defend claim (1). I offer an example that our intuitive judgments suggest is a case of understanding but in which justification appears absent. I also offer two
arguments for the following claim: to the extent that our concept understanding is unclear or vague, it is beneficial to sharpen it so as not to demand that understanding obtain in virtue of a thinker’s possessing justified beliefs. Given the inconsistency of the (1)-(3) above, establishing (1) shows that anything that supports either (2) or (3) provides reason to doubt the other. The exhibition of that tension is the major goal of this chapter. However, as it is still of interest to determine which of (2) or (3) should be rejected, in §3 I explore some costs and benefits associated with either strategy. The decision of which to drop ultimately rests upon broader epistemological considerations; but I suggest that a higher proportion of the evidence than one might expect favors (2) over (3). Thus, the present discussion provides one heretofore unnoticed reason to be suspicious of (3).

§1 What Understanding? What Knowledge?

In this section I argue that there is an important sense of understanding, distinct from (though possibly a species of) knowledge, that there is good reason to articulate more precisely. I also look at what typical objects of such understanding are, and how they are best compared to knowledge.

The word ‘understands’ is used in many different ways, including for purposes as diverse as hedging against uncertainty (see e.g. Elgin 34) and expressing sympathy. Sometimes, particularly when followed by a that-clause, it is clearly used interchangeably with ‘knows’, as in “I understand that you were late yesterday.” However, it is just as clear that ‘understands’ often appears in contexts where it cannot be replaced with ‘knows’ without change of meaning. “I know John” means something very different from
“I understand John”, and “I know the answer” at least sometimes means something different than “I understand the answer”. “I understand the liquid’s expansion” sounds felicitous, whereas “I know the liquid’s expansion” seems confused.\textsuperscript{96} Pointing out these differences in usage does not yet settle the question of whether understanding is a species of knowledge: for all that has been said, it is possible that claims of understanding \( \psi \) can be reduced to some more nuanced statement or statements of knowledge about \( \psi \).

One sense of ‘understanding’ denotes a particular cognitive achievement, distinct from knowledge. (For purposes of exposition I hereafter assume that there is exactly one such sense, but it suffices for defense of (1) to assume that there is at least one such sense.) This sense is important, both because it is a sense often explored\textsuperscript{97} and because it seems to be most tightly bound up with (successful) explanation. When explanations successfully elucidate some feature of the world, they result in the sort of understanding envisioned.\textsuperscript{98}

Paradigmatic instances of cognitive understanding seem to relate cognizers to states-of-affairs, facts, objects and theories.\textsuperscript{99} This is not a syntactic claim—syntactically, the word ‘understands’ functions almost (if not entirely) identically to the word ‘knows’.

\textsuperscript{96} For a more extensive discussion of the analogies and disanalogies between knowledge- and understanding-talk, see Kvanvig (forthcoming).
\textsuperscript{97} I take it to be the sense explored by e.g. Grimm (2006), Elgin (2007), and Kvanvig (2003).
\textsuperscript{98} It is largely because of this latter connection that arguing over the nature of understanding is not “merely verbal”, as described by Chalmers (2011). If we were to create new terms for understanding\textsubscript{1} that obtains in virtue of possessing justification and understanding\textsubscript{2} that does not, we would still have a substantive dispute over whether understanding\textsubscript{1} or understanding\textsubscript{2} is the type that are supposed to be an aim (or at least result) of good explanations (see §2.2 and Chapters 3 and 4 for more on this connection).
\textsuperscript{99} For present purposes I put aside questions pertaining to hermeneutic understanding and how it relates to the sort of cognitive understanding envisioned. All that I require is that we at least \textit{sometimes} understand things like theories and states-of-affairs in some sense other than merely understanding the words in which it is represented.
The claim is rather that the cognitive understanding relationship of interest tends to take objects other than propositions. That understanding qua cognitive achievement takes things (broadly construed) as its objects, whereas knowledge seems to most often be of propositions, does not preclude the possibility that understanding is still some species of knowledge. For example, it could be the case that understanding the state-of-affairs-such-that-φ-obtained (hereafter “state-of-affairs φ” or just “φ”) reduces to knowing a block of facts about φ. However, the disparity in apparent primary objects of understanding and knowledge suggests that it will likely prove no trivial task to specify precisely what sort of knowledge understanding might be.

To begin, I will focus on understanding of a particular state-of-affairs φ (such as the state-of-affairs-such-that-the-Challenger-exploded). For the sake of argument I will assume that understanding φ qua state-of-affairs requires understanding why φ obtained (hereafter why-φ).100 I will argue even on this expansive notion of what is required to understand φ, such understanding does not obtain in virtue of possessing justified beliefs about φ—in particular beliefs regarding why-φ or even that φ obtained (hereafter that-φ). Assuming that there are no other candidate types of beliefs about φ that understanding φ requires, I will conclude that one does not understand φ in virtue of possessing any particular justified beliefs about φ. Thus, to the extent that one is inclined to view such

100 Since I am assuming that understanding states-of-affairs involves understanding such propositions as why the state of affairs obtained, I can remain neutral regarding what precisely state of affairs are and in particular whether they have propositional structure. The argument is that even if understanding states-of-affairs requires understanding certain propositions, it still does not always require justified belief in those propositions.
understanding as a species of knowledge, one should want to deny that one always possesses knowledge in virtue of possessing justified belief.

§2 Justification—A Needlessly High Bar

In this section I argue in defense of (1), that one does not possess understanding in virtue of possessing justification. To be clear, the claim is not merely that one’s being in a particular state does not count as understanding in virtue of possessing justification, but rather the stronger claim that whatever state understanding is, one is not in that state in virtue of possessing justification. In §2.1 I give a case in which, intuitively, the subject understands φ, but not in virtue of possessing justification in his beliefs why-φ or even that-φ. So as not to be reliant on particular intuitions about extreme cases, in §2.2 I give two theory-driven arguments that, to the extent that our concept understanding does not dictate a clear answer to cases such as those in §2.1, we would do well to treat understanding as not obtaining in virtue of justification possession.

§2.1 Understanding, Justification, and the Challenger Explosion

In this section I defend (1). There are contexts in which we would attribute understanding of φ to someone despite that person apparently lacking a justified belief why-φ. In §2.2 and §3 we will explore the possibility that the subject’s beliefs in the following case are justified, but are justified in virtue of his possessing understanding, rather than conversely.

\[101\]
I assume that, whatever justification is, evidence of ~p undermines one’s justification in believing p. This is generally granted even by non-evidential theories of justification, such as reliabilism (at least on the assumption that the contrary evidence was itself acquired by a reliable mechanism. See e.g. Goldman (1979, p. 102)).

Consider the following case, which demonstrates that one can possess understanding while (apparently) lacking justification (Challenger):

Richard is an established and skilled scientist tasked with investigating the explosion of the Challenger space shuttle. His investigative instincts are excellent, and he is immediately suspicious of the O-rings’ ability to operate at cold temperatures. This inability of the O-rings is what actually caused the explosion. Sadly, before he goes public he is subjected to a deliberate cover-up, casting doubt on whether the O-Rings failed in this case. Arbitrarily large amounts of evidence can be presented that there really was no O-Ring failure, to the point that, whatever the bar is for justification, Richard does not have justification for his belief that the O-Rings caused the explosion. Nevertheless, Richard still understands the explosion. If he clings to his suspicions (which he might or might not retain with sufficient strength to qualify as beliefs) and continues acting like the O-Rings caused the explosion, it would seem that he still understands the state-of-affairs such that the Challenger exploded. Thus, one can understand a state-of-affairs not in virtue possessing justified belief why it obtained.¹⁰²

¹⁰² I am inclined to say that Richard understands when he maintains his beliefs, even after his holding those beliefs becomes irrational. However, if one thinks understanding is incompatible with doxastic irrationality, one can also envision that he no longer has sufficient credence to believe that the O-Rings caused the failure, but maintains suspicions that do not rise to the level of belief. For those who do not share the intuition in either case, see §2.2 for more theoretical reasons for attributing understanding without demanding justification.
There are two features of *Challenger* that need to be defended—that Richard possesses understanding, but that he lacks relevant justified beliefs. I will now defend each in turn.

Even if one shares the intuition that Richard understands, one might worry that this is not really evidence of Richard’s *actually* understanding but only that we have good pragmatic grounds to *treat* Richard as understanding. Hawthorne (2000 p. 202-204) has argued that there are cases (e.g. where one really has to get to Boston) where we find it pragmatically felicitous to refer to someone’s true belief (e.g. that the way to get to Boston is θ) as “knowledge”, even though it does not literally qualify—perhaps the attribution of “understanding” to Richard is similarly pragmatically driven. However, the knowledge attribution in the case Hawthorne envisions and our understanding attribution in response to *Challenger* are not analogous, as in the latter the exact context has been so underspecified. I have said nothing about what my goals would be in asking Richard about the explosion, and so there is no basis for thinking the attribution of understanding depends on some particular pragmatic factor.

One might also worry that Richard possesses some limited form of understanding, but less than complete understanding. But this is no concern at all, for it acknowledges that some understanding can be possessed in the absence of relevant justified belief.

As further evidence that Richard understands this explosion, it is worth noting that other experts, who have been spared the disinformation campaign to which he has been subjected, will defer to him regarding this very explosion.
It remains to show that Richard does not possess any relevant justified beliefs about φ. He certainly possesses some general justified beliefs. He has justified beliefs regarding what explosions are, what the word ‘explosions’ means, and (to the extent that they are generally known) what physical laws govern exploding space shuttles. He might even have interesting counterfactual justified beliefs to the effect that if the O-Rings had failed, they would have caused the explosion. Nevertheless, none of those justified beliefs seem about this explosion in particular. In contrast to what was said with respect to understanding, one should not plead here that Richard “sort of” possesses a justified belief. While there is room for considerable variation with regard to height of the bar for determining what suffices for justification, in Challenger Richard possesses arbitrarily little justification—that should put it beneath any bar, however determined.

Finally, one might counter that, while Richard lacks knowledge of why-φ, he might possess some other justified belief about φ in virtue of which he understands. There are two problems with this approach. First, it is not at all clear what those other justified beliefs would be, or how they could in any combination amount to understanding of φ. More problematically, one can generate Challenger-style cases for any belief about φ. Consider Challenger*, which is just like Challenger, with the additional twist that Richard has been fed misleading information to the effect that no explosion occurred. Yet again he retains his suspicions that there was an explosion caused by O-Ring failure, and once again he continues to act as if those suspicions were correct. While intuitions about this modified case are perhaps more ambivalent, I contend
that it is a case where he understands while apparently lacking any justified beliefs pertaining to $\varphi$.

Richard possesses understanding of $\varphi$ in the apparent absence of any justified beliefs about $\varphi$. And it would be extremely peculiar if understanding this explosion could be reduced to knowledge made up of justified beliefs that are not about this explosion at all. Thus, one does not understand $\varphi$ in virtue of possessing justified beliefs about $\varphi$.

§2.2 Theory-Driven Reasons to Treat Understanding before Justification

In this sub-section I present two arguments in favor of affirming (1), that understanding does not obtain in virtue of a thinker’s possessing justification.

The first argument is that on many if not most independently motivated accounts of understanding, Richard understands. While all the theories in question impose some veridicality constraint on understanding, there is no reason to think such constraints line up with the demand that the beliefs underlying the understanding be justified. Typically, such accounts will measure understanding in terms of some first-order success, which, unlike justification, can be retained in the presence of arbitrarily large amounts of contrary evidence. Richard possesses the ability to make qualitative calculations about the explosion, as required by De Regt and Dieks (2005); De Regt & Dieks require that such calculations be accurate, but Richard’s calculations are accurate—he just lacks justification for believing in them. Richard also has an accurate model of the explosion (as required by, e.g., Knuutilla & Merz (2009)), and represents the explosion in ways that enable him to make effective inferences (as required by the account presented in Chapter...
5). Even on Jay Rosenberg’s (1981) irrealist theory of understanding, Richard possesses understanding in virtue of the fact that others in the know treat him as understanding. For the purposes of this chapter I would like to remain neutral regarding the proper account of understanding, but it is noteworthy that several independently motivated accounts (including the one defended in Chapter 5) support what I take to be the intuitive results in with respect to *Challenger*. There is simply no need to invoke justification to account for the general data regarding when and why we treat certain people as understanding and others as not, and so no need to require justification (even indirectly) when building a theory of understanding.

The second argument in favor of the claim that one does not understand in virtue of possessing justified beliefs abstracts away from particular accounts of understanding and focuses instead on what role we want the concept *understanding* to play. If it turns out that one sense of (the concept) *understanding* is particularly theoretically fruitful, that gives us reason to hone our concept along those lines, or at least to refocus some attention on that sense. I argue that there is such a reason to support treating the concept of *understanding* (and hence, in some contexts, understanding itself) as being applicable to subjects not in virtue of their beliefs being justified.

As noted in §1, the sort of understanding in which this paper is interested is that which is most directly tied to explanations. It has been fairly universally acknowledged that, at least among the other things they do, explanations create understanding (see e.g.

---

103 For a defense of this general strategy with respect to (the concept) *knowledge*, see Craig (1999).
Michael Friedman (1974), Jaegwon Kim (1992), Paul Thagard,\textsuperscript{104} and Hempel (1965))

Some have even argued that it is in virtue of something’s propensity to cause understanding that it counts as an explanation in the first place (Scriven (1962)). I argue that we would not expect the product of explanation to require independent justification.

The central point is that one important way explanatory statements \textit{gain} justification\textsuperscript{105} is by being inferred as the best possible explanation among the viable candidates. That is to say that certain explanatory statements gain justification in virtue of their being, if true, the “best” explanations. There is, sadly, no consensus on what makes an explanation the “best” to infer. One intuitive heuristic is that one of the things that contributes to the quality of an explanation is the extent to which it is conducive to understanding. In the most thorough investigation of the subject, Peter Lipton (2004) suggests that we choose from among candidate potential explanations the “loveliest”, which he also characterizes as the one that would “provide the most understanding” (Chapter 4.0). It would thus be theoretically fruitful to be able to appeal to the understanding engendered by an explanation when determining if that explanation should be inferred. If, however, understanding $\varphi$ already depended on one’s relevant beliefs about $\varphi$ be justified, then it would be difficult to use IBE to learn new things (i.e. gain

\textsuperscript{104} Personal correspondence.

\textsuperscript{105} As elsewhere in the paper, ‘justification’ is here used to denote a property of a belief as held by a particular person or group of persons, as distinct from a more objective sense, as for example a relation between a body of evidence and a proposition. However, whether a belief is justified for a person is being assessed independently of that person’s actually basing her belief on that justification.
new justifications pertaining to beliefs) about φ. One could deny that IBE is justification conferring, but given its sheer ubiquity\textsuperscript{106} this would be a hugely revisionary strategy.\textsuperscript{107}

The problem IBE poses for a demand that understanding be justified is best illustrated with an example. Suppose Copernicus had just espoused his heliocentric worldview. At the time, Ptolemaic astronomy was at least as predictively successful as Copernican, yet there was some force to accept the latter. One intuitive way to explain the advantage of Copernican astronomy is that, if it were true, coming to believe it would better enable us to understand the motion of the planets. However, if understanding the motion of the planets required justified beliefs about that motion, the justification could not have been present before the inference was made. If it had been, there would have been no need to infer\textsuperscript{108} that the motion was as Copernicus described—we would already possess understanding underwritten by justified belief that the solar system was heliocentric. Thus, when surveying possible understanding states, we must be able to

\textsuperscript{106}In addition to Lipton (2004), see Harman (1965) which claims that IBE underlies many other inferential forms, such as enumerative induction.

\textsuperscript{107}The conclusion of the argument from IBE is admittedly weaker than (1), as one can only engage in IBE if one already has some justified beliefs about the occurrence of the explanandum (thing to be explained). Thus in a sense the understanding to which one appeals when engaged in IBE does obtain in virtue of possessing some justified belief about the phenomenon under discussion (i.e. that it occurred). Nevertheless, that sort of justified belief, while about φ, is somewhat ancillary to the understanding itself. This is best demonstrated by the fact that the beliefs that are the most credible candidates for constituting the understanding—e.g. the belief that the cause of the explosion was that X, or the belief that the reason that happened was that Y—need not be antecedently justified. Rather, some other beliefs—not those making up the understanding itself—about the understood must be justified. Thus, while the understanding appealed to by IBE depends on the possession of some justified beliefs about the understood, they are not the beliefs the proponent of (2) and (3) should want to claim ground the knowledge of which understanding is a species. Thus a weakened version of (1)—that understanding does not always obtain in virtue of being composed of justified beliefs—is sufficient to at least demonstrate a strong tension between (2) and (3).

\textsuperscript{108}Importantly, as already alluded to above, the inferential process is here being taken as justification conferring. This is independent of the psychology of particular inference makers. See n. 105)
conceptualize such states in the absence of justification.\textsuperscript{109} It would thus be theoretically fruitful to have a concept of understanding that did not force us to immediately demand justification of the beliefs (or perhaps other mental states) underlying that understanding.

One could object that the understanding surveyed when engaged in an inference to the best explanation is only purported understanding. By analogy, since something can only be an explanation if it is true, before we decide on a best explanation we need to survey potential explanations. This fact does not by itself imply that explanations do not require truth. One might think that we could make a similar move regarding the surveyed understanding, and argue that while understanding generally requires justification, we need only appeal to purported understanding when engaged in IBE. This move would, however, still place the (purported) understanding explanatorily prior to the justification. While understanding, as opposed to purported understanding, might always be present only with justification, it is the beliefs’ (or other representational states’) qualifying as purported understanding that explains their being properly inferred, and hence justified.\textsuperscript{110} There is no (plausible) corresponding move regarding the priority of explaining over truth, as it is not in virtue of something’s being an explanation that it is true. The asymmetry between talk of potential explanations and purported understanding stems from the fact that an explanation’s quality is merely evidence for its truth, which truth would have obtained or failed to obtain regardless of the quality of the explanation.

\textsuperscript{109} One might note that in the future the understanding imagined will involve only justified states, because they will be justified by the very inference that we are making; this would still not be a case where understanding obtained in virtue of a state’s possessing justification, but rather one where justification obtained (at least in large part) in virtue of that state’s propensity to cause understanding.
\textsuperscript{110} Similarly, Williamson (whose view is discussed in §3.2, especially n. 118) argues that one’s claims are justified in virtue of their relation to knowledge, and not conversely.
The justification of the (beliefs or other mental states comprising the) understanding, by contrast, is at least partially \textit{constituted} by the presence of the quality of the understanding. Consider the following analogy: footprints on a beach provide evidence that someone was walking there. But the walker is independent of that which provides evidence for it, and so we can make sense of both explanations and walkers independently of each other. By contrast, if a choice’s being the best is constituted by it causing the most net well-being, then we would do well (if possible) to have a concept of well-being that does not depend on such states’ choice-worthiness.

Clearly, having such a justification-free concept of understanding is not the only way to make sense of IBE, but the fact that it would enable us to model IBE is one reason to favor having it. This reason is wholly independent of the reason provided by intuitions regarding cases like \textit{Challenger}, and the independently motivated theories of understanding that affirm those intuitions.

§3 (2) or (3)?

If (1) is established, then either (2) or (3) must be false. That is, either understanding is not a species of knowledge (not-2) or not all knowledge one possesses obtains in virtue of one possessing justification (not-3). In this section I explore whether the present problem favors one approach over the other. In §3.1 I argue that, while there is no airtight argument in favor of (2), there are enough inductive arguments to jointly give (2) some plausibility. Moreover, one well-known independent objection against (2) does not pose any serious difficulties. In §3.2 I look at (3), and argue that while there is
also a great deal of inductive evidence in support of (3), the trend of some recent arguments might lead us to call (3) into question. I thus conclude tentatively that the truth of (1) gives us more reason to doubt (3) than (2).

§3.1 Is Understanding a Species of Knowledge?

In this subsection I look at 4 arguments that support the claim that understanding is a species of knowledge, in increasing order of persuasiveness. I then look at one objection and argue that it is confused. I thus conclude that there is a strong inductive case for believing that understanding is a species of knowledge.

Stephen Grimm (2006, p. 522) argues that understanding is subject to an anti-luck condition, which would be best explained by seeing such a condition as a token of the more general anti-luck condition applicable to all knowledge. He imagines a situation wherein a person walks by a blacksmith apparently smashing chestnuts with a hammer. Unbeknownst to our spectator, the vast majority of the time, right when the blacksmith is about to strike the chestnut, it explodes as a result of the heat of the anvil. The spectator, not knowing this, assumes that the chestnut she saw explode did so because of the hammer blow, and further believes that this knowledge entails that she understands why the chestnut exploded. Oddly enough, on this one occasion the blacksmith forgot to heat up the anvil, and so the hammer blow actually did cause the explosion. Still, Grimm asserts, "the answer seems to be" (521) that she does not understand the explosion. The problem with this case is that the intuition to which it appeals is not universally shared.
Jonathan Kvanvig (2003, p. 197-8) elsewhere says of a structurally analogous case\footnote{111} that it would be one of genuine understanding. While Grimm might well be correct regarding the cases in question, we would do well to find an argument that understanding is a species of knowledge that relies less on controversial intuitions.\footnote{112}

The second argument in support of the claim that understanding is a species of knowledge is that, on our best theories of explanation, explaining is typically a matter of putting forward certain propositions—it is thus natural to think that the understanding that is effected by such explanations requires knowledge of those propositions. It is right to think that explanations do bear an important causal connection to understanding, which I discuss in depth elsewhere.\footnote{113} However, in order to draw the conclusion that understanding is a species of knowledge from the truisms that explanations have propositional content and that understanding is typically caused by explanations, one needs an additional premise to the effect that the only way to gain understanding from an explanation is by coming to know the propositional content the explanation presents. This premise—that one must bear the particular knows-that relation to an explanation in order to gain understanding from it—is independent of the details of any particular account of explanation. Put differently, theories of explanation are (typically) silent regarding how to uptake the explanations they identify. If, for example, one thinks that James Woodward’s interventionist causal account of explanation is entirely correct, then one

\footnote{111}{Grimm denies that the cases are really analogous; assessing this claim is beyond the scope of the paper. If he is correct, that only gives more reason to favor retaining (2) over (3).}

\footnote{112}{In defense of Grimm’s intuition, some accounts of understanding (including the one presented in Chapter 5) do require that understanders be able to counterfactually reason well about the understood object, which feature seems absent in this case.}

\footnote{113}{See Chapter 3.}
would likely be forced to maintain that understanding involves some relation between the understander and facts about causation—but which particular relation one must bear to these facts is wholly underdetermined by the account of explanation itself. Gleaning understanding from an explanation could be a result of knowing the proposition that X causes Y, but, for all Woodward’s theory of explanation or the explanation itself tell us, it could also result from one merely believing it. This point generalizes to other theories of understanding derived entirely from corresponding theories of explanation. Still, while the theories of explanation do not demand that the content of an explanation be known in order to be efficacious, knowing is a plausible candidate for the relation one must be in vis-a-vis explanatory content in order to possess understanding. It is hard to imagine that mere blind-belief in an explanation confers understanding; there is thus at least a burden on one who claims explanations can effect understanding without being known to say what particular epistemic relation one must bear to the explanation for it to do so.

The third reason to suspect that understanding is a species of knowledge, also noted by Grimm (2006), is simply that understanding being a species of knowledge is the accepted view in the philosophy of science. Given that one of our main interests in exploring understanding is its link to explanation, which topic has for decades been largely in the purview of the philosophy of science, the views of philosophers of science should not be treated lightly. At first glance this argument from authority might not seem persuasive, as with all such arguments one might well wonder why we do not simply

---

114 Even the unificationist views of Michael Friedman (1974) and Philip Kitcher (1989), which insist that something is explanatory in virtue of the effect it has on one’s total store of knowledge, fall short of demanding that the explanation itself can only have this effect if known.
survey the reasons of the alleged authority directly. However, given the extent to which philosophers of science often engage with actual scientific practice, facts about what locutions they have come to find natural are revealing. Their treatment of understanding as a species of knowledge suggests, in a way not subject to direct argumentative analysis, that such is common practice amongst practicing scientists.

The fourth, and most impressive, defense of the claim that understanding $\phi$ requires knowledge of $\phi$ stems from basic linguistic intuitions. For example it does sound odd to say “I understand the Challenger explosion, but don’t know that the Challenger exploded”, and this oddity remains with respect to other objects of understanding and knowledge. “I understand Joe but I don’t know Joe” and “I understand (proposition) $p$ but I don’t know $p$” sound equally perplexing. The most obvious explanation of this infelicity is that understanding just is a species of knowledge. This is evidence in favor of trying to maintain the link between understanding and knowledge against whatever objections may come.\footnote{Nevertheless, this evidence is not quite as decisive as it might at first appear. It does not really show that $possessing$ understanding of $\phi$ requires $possessing$ knowledge that $\phi$, but rather just that one cannot felicitously $assert$ that one understands $\phi$ without knowing that $\phi$. However, in the absence of a further pragmatic story regarding why such assertions would be infelicitous, we should take it as a (defeasible) assumption that the sorts of sentences in question are semantically ill-formed. This in turn suggests strongly that understanding is a species of knowledge.}

I now examine one argument against understanding being a species of knowledge, and argue that it is not up to the task. Catherine Elgin contends that the knowledge and understanding relations take different relata. Where knowledge is a relation between
individuals and propositions, understanding is a relation between individuals and whole theories. Moreover, understanding does not so much as depend on knowledge even of whole theories. This denial of propositional understanding, however, rests largely on a confusion. Elgin defends taking whole theories as the objects of understanding as follows:

The epistemological standing of 'Athens defeated Persia in the battle of Marathon' then derives from its place in a more comprehensive understanding of the history of Greece. That is, the proposition derives its epistemological status from a suitably unified, integrated, coherent body of information. This is the core conception of understanding. (Elgin 2007 p. 35).

She goes on to assume that "Understanding is primarily a cognitive relation to a fairly comprehensive, coherent body of information." (Elgin 2007 p. 36) But it does not follow from the fact that one’s understanding depends on possession of a “suitably unified, integrated, coherent body of information” that one's understanding takes as its primary object only whole bodies of information. One's knowledge that a particle is going through the cloud chamber depends on a host of theoretical knowledge, but that does not prove that such knowledge is not primarily of a proposition. Perhaps when Elgin says that understanding "derives" form coherent bodies of information she means to invoke something stronger than mere dependence, but if so the claim the proposition derives its status from the entire body of knowledge is under-supported by her example. (For all that

116 Elgin (2007, p. 36) explains why theories should not, for these purposes, be treated as just one long conjunction. To summarize, understanding of a theory seems to depend on understanding some of its critical theorems, without necessarily understanding all of them.
has been said, Elgin might be right to have stressed that possessing understanding depends on some holistic feature of one’s epistemic system.)

The state of play thus far is one where there is some reason, albeit not dispositive, to affirm that understanding φ is a species of knowledge about φ, and little independent reason to deny this claim. In the next sub-section I examine some of the costs and benefits of denying that one possesses knowledge in virtue of possessing justification.

§3.2 Knowledge without Justification

The inductive argument for the claim that all of one’s knowledge is possessed in virtue of possessing justification is strong: The theory (JTB) that one knows ψ if and only if one possesses a justified true belief that ψ was the standard analysis until recently, and seems to be extensionally adequate in most common cases. Though Edmund Gettier (1963) found counterexamples to the sufficiency of JTB as an account of knowledge, this only immediately suggests adding conditions to what is required for knowledge, rather than subtracting any.

In addition, Timothy Williamson (2007) explicitly defends the necessity of justification. He argues that if our evaluations of justification are normative, then saying that there might be unjustified knowledge is saying that someone can be criticized for believing what they know. But that would mean that there are cases where we ought to give up something which we know, which is “silly” (Williamson 2007, p. 111). Moreover, if knowledge and justification can come apart, then it is difficult to know what to say about a belief that is the product of a valid inference from one piece of unjustified
knowledge and one justified belief that is not knowledge. Such an inferred belief seems like it should have some positive epistemic status, but it is neither knowledge nor justified. Notice however that, even if Williamson is correct,\textsuperscript{117} that shows only that there cannot be unjustified knowledge; one could grant that while still maintaining that one does not possess knowledge \textit{in virtue} of possessing justified belief.

Despite the long pedigree of views that one possesses knowledge only in virtue of possessing justification, some recent approaches have eschewed JTB-like views in favor of a radically different approach. Williamson himself—motivated in large part by the ability of post-Gettier epistemologists to produce counterexamples to the sufficiency of any proposed analysis of knowledge—proposes that knowledge is an unanalyzable state. Specifically, it is the most general factive mental state; that is, if one is in any other factive mental state in relation to proposition p—e.g. seeing that p or realizing that p—one also knows that p (Williamson Chapter 1.4). According to Williamson’s view, then, knowledge possession does not obtain \textit{in virtue} of possession of justification.\textsuperscript{118} On this

\textsuperscript{117} A full response to Williamson (2007) is well beyond the scope of this essay, but I am not sure either consequence he draws out from denying that knowledge requires justification is as dire as it may at first appear. Little argument is given to the conclusion that it is always silly to give up on knowledge—one can certainly imagine cases where it would be pragmatically advisable to do so, so the idea that it might sometimes be rationally advisable might not be as outlandish as it might first seem. Regarding beliefs that are the product of inferences from unjustified knowledge and justified beliefs that are not knowledge, it is not clear to me that we should expect them to have \textit{any} positive epistemic feature. The offspring of two beliefs each of which have a positive and a negative feature might well have inherited the worst of both; for all the information we have the conclusion of such an argument can be both unjustified and unknown. These points are not intended to be a full argument against Williamson’s point which, as indicated in the text, one can accept while still maintaining that knowledge does not obtain in virtue of the possession of justified belief.

\textsuperscript{118} The important point for my purposes is the point about the respective explanatory priority of knowledge over justification. Williamson also contends that one’s evidence is simply one’s store of knowledge, in which case all of one’s knowledge is trivially justified. This latter part of his view is separable from the general approach to knowledge and, in any case, discards the problematic implication that one possesses knowledge \textit{in virtue} of a belief’s being justified.
conception of knowledge, it is possible that Richard knows that and why the *Challenger* exploded after all. In virtue of understanding the state-of-affairs such that the Challenger exploded, Richard understands that the Challenger exploded and why the Challenger exploded. However, one can only understand that-χ or why-χ if χ is true—thus, in virtue of his understanding, Richard is in a mental state with regard to χ that he can only be in if χ. Therefore, Richard is in a factive mental state with respect to χ and so must, on Williamson’s account, knows that χ. He knows that and why it happened *in virtue of the fact* that he understands that and why it happened, in the same way that my 17 month-old can know there are cheerios in front of her in virtue of the fact that she sees that there are cheerios in front of her. Richard and my daughter both stand in an epistemic relation to the known fact—understanding the explosion and seeing the cheerios, respectively—that one could only stand in if the fact obtained. Therefore, they know it. This suggests that we should reject the claim that one knows in virtue of possessing justified belief (3) in favor of the claim that understanding is a species of knowledge (2).

A related conception of knowledge that does not demand antecedent justification can be found in the literature on knowing-how. For example, Berit Brogaard argues that “some knowledge-how states are pairs of a representational informational state and a bodily ability state.” (2012, p. 154) Specifically, in order to know how to χ one must "at least at some level of information processing," (ibid.) represent some series of steps θ as
being a way to \( \chi \). If such bits of information are veridical then their possessor could possess knowledge, even in the absence of justified belief.

Brogaard’s treatment of knowledge-how suggests a possible parallel treatment of understanding. Where there is an intuitive connection between knowledge-how to \( \chi \) and ability to \( \chi \), there is a similar connection between understanding \( \varphi \) and the ability to reason well regarding \( \varphi \). Thus, where knowledge-how to \( \chi \) consists in a veridical representation state of some way to \( \chi \) wedded to an ability, one might suppose that understanding \( \varphi \) consists in a veridical representation state of (contextually salient features of) \( \varphi \) along with an ability to make reasonable inferences regarding \( \varphi \). Call this theory B*U, for Brogaard* Understanding. Clearly, more needs to be said about what representation states are, in virtue of what they count as veridical, and what role context plays in determining which aspects of \( \varphi \) are relevant. However, none of these problems are unique to the B*U, and so we may reasonably hope that they will not prove insolvable. This theory of understanding does not require antecedent justification of any claim about the understood, and is thus largely immune from the concerns raised by Challenger.

One notable aspect of B*U is that it is almost identical to the theory of understanding I motivate on independent grounds in Chapter 5. Recall that on that account, understanding a state-of-affairs \( \varphi \) requires representing \( \varphi \) in such a way that one could (in some contextually relevant set of counterfactuals) successfully reason about \( \varphi \).

---

\[119\] It is not clear what the relevant “level of information processing” is. One possible answer would be that it is the lowest level at which the information has the potential to be directly consciously accessed—this would eliminate a possible concern that knowledge-how would be under-intellectualized to the point that we would know how to calculate depth perception.
If one expects that such successful reasoning could only occur if the representation is basically veridical, then understanding on the view I defend involves the pairing of a veridical representation with an ability to reason about the understood. But veridical representation paired with ability is what Brogaard says knowledge-how amounts to. The account of understanding in question then amounts to an extension of Brogaard’s treatment of knowledge-how to encompass theoretical and inferential abilities, which is precisely the relation one might expect understanding to bear to knowledge-how—understanding φ is, to some extent, knowing how to think about φ. If one is inclined to retain the view that understanding is a species of knowledge, Brogaard's account of knowledge and my own account of understanding are thus mutually reinforcing (at the expense of (3)).

§4 Conclusion

In this chapter I have argued that one does not possess understanding in virtue of possessing justified beliefs. Thus it cannot both be the case that understanding is a species of knowledge and that one possesses knowledge in virtue of possessing justification. There is strong inductive evidence in favor of considering understanding as a species of knowledge; there are also conceptions of knowledge that both deny the traditional claim that one knows in virtue of possessing justified belief and bear some resemblance to independently motivated accounts of understanding. These facts, taken together, give some reason to favor denying that one knows in virtue of possessing justified beliefs. Ultimate resolution of which statement of the inconsistent triad should
be rejected will hinge in large part on independent considerations in the study of knowledge, though the present result might inform such studies with respect to a potentially unrealized cost/benefit of particular theories.
Chapter 7: Explanatory Asymmetry and Pragmatic Explanation

Among the traditional litmus cases for a theory of explanation, perhaps none has been more historically important than those involving explanatory asymmetries. It is claimed that we explain later events by citing earlier ones, and not conversely; it is analogously claimed that we explain effects in terms of causes, but not conversely. This has been thought to be a decisive objection against the classic D-N account of explanation, the unificationist account of explanation, and pragmatic explanation, all for the same reason: since none of these accounts build in criteria favoring past/causal states over future/effect states, the tools they provide to identify explanations pick out the bad states as well as the good ones. Among the popular contenders, only the causal account of explanation was thought able to handle these cases.

These counterexamples are serious. There are, however, pragmatic accounts that have the resources to treat such examples, not just by classifying them correctly but by explaining their very recalcitrance. The examples thus serve as further evidence in support of (at least some) pragmatic accounts.

---

120 On some such accounts what are picked out are statements of the states’ obtaining, while on others what explanations pick are the states themselves. This question is orthogonal to the one under discussion, and will not be explored here further.
In the next section I will discuss the two types of explanatory asymmetry and why they are thought to pose insurmountable difficulties for D-N and unificationist accounts of explanation. In §2 I discuss an attempt by Bas Van Fraassen to deal with the counterexamples, to the effect that explaining causes in terms of effects and/or earlier states in terms of later states is sometimes acceptable. Van Fraassen's argument for this position takes the form of an example which, regrettably, does not prove what he says it does. Nevertheless, in §3 the torch is taken up, and it is argued that a modified version of Van Fraassen's strategy, centered on a different example, shows that explanatory asymmetries can be reversed in some contexts. I argue that certain pragmatic views will have in their appeals to contextual variance the tools necessary to account for our intuitions regarding explanatory asymmetries. They will do so in part by taking notice of the oft underemphasized fact that there are different (generally pragmatic) constraints on what sort of explanations will be acceptable in different contexts. In §4 I then turn to causal accounts of explanation and argue that the sorts of tools causal accounts utilize to block bad cases of later facts explaining earlier ones will also likely block the newly presented good cases of later events explaining earlier ones.

§1 Explanatory Asymmetries

Suppose we have an argument whose premises are statements of the laws of light propagation, the angle of the sun relative to the surface of the Earth, and a statement that a particular flagpole on that surface is of a certain height; the conclusion is that the
flagpole's shadow is of a particular length.\footnote{As already noted in n. 7, we would actually need more information than this. We would need to know about the curvature of the Earth and, more importantly, the location of any other light sources that might strike the shadow. The tradition in the literature is to ignore such details; it is not clear that they change anything about the substance of the example. It should be admitted that Neil Tennant (2008) has argued that in at least one famous case of explanatory asymmetry—the case of explaining a pendulum's period by citing its length—the details actually do matter. There is no reason to suppose that his results generalize to all other such examples, however.} We have, according to the proponent of the Deductive-Nomological account of explanation, explained the shadow's length. Call this explanation Real-Shadow. But there is a catch—if we simply swap the statements of the flagpole's height and the shadow's length, we have a structurally identical argument. Call the newly minted “explanation” of the flagpole’s height in terms of its shadow Pseudo-Shadow. If D-N counts Real-Shadow as an explanation of the shadow's length, then it should also count Pseudo-Shadow as an explanation of the flagpole's height. But this is (supposed to be) absurd. The absurdity is that the shadow's length does not tell us why the flagpole is the height that it is, whereas the flagpole's height does tell us why the shadow is the length that it is. The relation between flagpole heights and shadow lengths is asymmetric—the former explains the latter, but not conversely.

A second well-known example of explanatory asymmetry concerns the derivation of an eclipse from other facts about the state of the system. If one derives from the positions of the moon, the Earth, and the sun on July 18th, 2012 and the laws governing their motion that there will be an eclipse on November 13th, 2012, that seems like a pretty good explanation of why the sun will be blocked that day. Call this explanation Real-Eclipse. But the laws of motion for large objects are time-symmetric, and one can just easily derive the necessity of the eclipse from those laws and information about the sun,
moon and Earth on New Year's Day 2013. Call this “explanation” Pseudo-Eclipse.
Pseudo-Eclipse, however, seems to be no explanation at all. Though earlier states of a system and principles of change seem to explain later stages of that system, the converse is not true. The relation of early states and later states of the same system is asymmetric.

Do the examples of the flagpole and the eclipse illustrate one sort of asymmetry or two? While there is some similarity between the cases, they are also importantly different. A few quirks aside, the relevant fact in the shadow cases seems to be that the flagpole's height *causes* the shadow's length, whereas the length does not cause the height. It is not clear, however, that the best way to think of the relation between earlier and later states of a physical system is in causal terms. One interesting feature of Real-Eclipse is that the time-slice of the system being cited in the explanans is not necessarily one where there is any *change* in the evolution of the system. Though the details vary with one's chosen account of causation, causes are generally thought of as change-relating. A perfectly stable glass window pane is caused to shatter by a baseball hitting it, or by a freak quantum occurrence, or even by a flaw having been introduced to the structure at the time of manufacture—in general, because something happened to it. It sounds odd to say that it was *caused* to break by the natural evolution of the system (as distinct from the slightly less awkward claim about reasons that it broke *because* of the natural evolution of the system).

---

122 The first quirk is that the alleged derivation does not actually provide enough information (see n. 121). Furthermore, the choice of explananda is a bit unusual—since shadows are absences, a metaphysician with a particularly austere ontology might think that what needs to be explained is not the shadow but the pattern of light surrounding it. I will follow the tradition in this literature of ignoring those details.
Are examples of causal asymmetry reducible to examples of temporal asymmetry? They of course will be if, in fact, all causes precede their effects. Outside of unusual examples in advanced physics, it seems that most of what we know does support the claim that causal asymmetry is just a species of temporal asymmetry, albeit a very important one. Nothing that follows turns on this point, however.

Granted that examples of explanatory asymmetry are problematic for D-N, one might suppose that theoretical descendants of D-N, built with such examples in mind, would fare better. As a general rule though, most do not. As James Woodward argues (2003, Chapter 8), requiring that explanations be unifying does not do the job of preventing effects or later events from being ruled as explanatory of causes or earlier events. Admittedly in this world there are more opaque objects than shadows, so learning to derive the length of the shadow from the height of the flagpole does yield some cognitive efficiency. But we can imagine a world where every object with a height also casts a shadow, in which case learning either direction of the derivation would equally well enable us to systematize our information about heights and shadows. Indeed if there were more shadows than objects with heights, it would be more efficient to learn the properties of shadows than the properties of flagpoles. The problem for unificationism is not that it grounds explanatory asymmetry in contingent facts—since our explanatory practice is constructed for the actual world, appealing to features of the actual world to explain aspects of that practice is a reasonable move. The problem with unificationism is that it appeals to the wrong contingent facts. The intuition is that even if we were in a world where a small number of flagpoles cast a large number of shadows, the flagpoles
would still figure into the explanations of the shadows just as much as they do in the actual world, and not conversely. Thus that contingent fact that unificationist theories cite to explain why we privilege causes is not what actually accounts for that preference.

The examples thus suggest a shift from D-N-type views to a new account of explanation. The most obvious alternatives are causal accounts, which block asymmetric explanations by building respect for actual temporal and causal asymmetry directly into the theory. I will explore a different possibility—that temporal and causal favoritism is simply the result of earlier states being pragmatically favored in most contexts (based on our interest in predicting and influencing future events). If that were correct, one would expect that, in some contexts, the typical direction would be reversed and we could use later and/or effect states to explain earlier and/or causal states. In the next section I present one attempt by Bas Van Fraassen to construct such a case, and argue that it does not work. However, such cases exist. In §3 I discuss one such case, and conclude that asymmetry can be explained by the preferences of speakers and audiences in most explanatory contexts. Having established a viable alternative to causal views in dealing with asymmetry, in §4 I turn back to those causal accounts.

§2 Van Fraassen’s Tower

Bas Van Fraassen (1980) argues that what counts as a good explanation varies contextually, and that it is only usually the case that context dictates that causes explain effects and earlier events explain later. He purports to give a context where the
asymmetry is reversed. If successful, he will have shown that the very relation explanans
must have to their explanandum varies with context, and so gone a long way towards
defending a pragmatic account of explanation.

The scenario Van Fraassen (1980, 132-4) describes is (narrative flourishes aside)
one in which a specific tower was built just so that its shadow would reach a certain
point. We are supposed to conclude that the shadow figures in the explanation of the
tower, and that any preference for explaining shadows in terms of towers (and flagpoles)
is wholly contextually determined. The problem is that, in this scenario, it is not at all
obvious that the effect-state actually figures into the explanation of the cause-state. There
is no doubt that the builder’s intention to have the shadow reach a certain point is part of
the explanation of the tower’s height, but that intention can in turn be fully explained
without appeal to the future state of the tower actually casting a shadow that reaches that
point. If the tower had been completed at night and unexpectedly destroyed before
morning, it would still have had the same height, despite never actually having cast the
envisioned shadow. The shadow length itself thus appears explanatorily inert.

§3 A Simple Example

In this section I give a straightforward case where a later fact apparently figures
into an explanation of an earlier fact. I argue that this example is of a quite common type.
Suppose Johnny is being asked to write a report on the assassination of Archduke
Ferdinand, explaining its importance. In the report, no doubt part of the explanation
would have to discuss the start of World War I—the report would be unlikely to garner a good grade without including such discussion. But World War I began after the assassination. In this case one is genuinely explaining an earlier event in terms of a later one. Call the explanation in Johnny’s paper ARCHDUKE. It is difficult to spell out ARCHDUKE explicitly without committing to a particular account of explanation, as one’s account affects the very form any explanation must take (e.g. argument or system of equations). As a rough approximation, in statement form it is:

ARCHDUKE: The fact that World War I resulted from the assassination of Archduke Ferdinand explains the fact that the assassination of Archduke Ferdinand was important.

One might object that the importance of the assassination is not really a feature of the earlier state-of-affairs, but only arises later. As evidence for this, one might cite the fact that, had the world ended immediately after the assassination, it would not have been important (on a global scale). While this is true, it only shows that the importance depends on later events, not that the importance, viewed from an atemporal perspective, arises only later. Intuitively, the importance of the assassination of the Archduke is a feature of the assassination of the Archduke. And examples akin to ARCHDUKE are perfectly common—whenever we explain the importance of something or the meaning of something we are just as likely to need to invoke later states to explain earlier ones as conversely.
The critical feature of ARCHDUKE is that the explanandum is a fact which bears a relational, explanatory property with a later fact. In principle, there is nothing about the example that turns on its normative or evaluative dimension. Nevertheless, it is surprisingly difficult to find an example from the physical sciences where a later fact explains an earlier one, without invoking some normative notion.

It is perhaps the difficulty of constructing a non-evaluative/non-normative example of a later fact explaining an earlier one that led Van Frassen astray. When seeking a counterexample to the alleged impossibility of effects explaining causes, he tried to construct a story where the answer to the question of why the tower had a certain height invoked the shadow. But, at least on the most straightforward reading of this question, there is no such answer—the very phrasing of the request for explanation in terms of the question of why the tower had a certain height (presumably rather than another) narrowed the explanandum to the simple fact that the tower had that height; it is possible that the only thing that will ever satisfactorily explain the obtaining of a simple fact will be appeal to earlier facts/causes. Had Van Fraassen instead constructed the story as one where what was being explained was the meaning of the tower’s being that height, he would not have run into such difficulties.

Van Fraassen has the formal resources to construct an example that evidences the desired reversal of standard asymmetries. On Van Fraassen’s theory of explanations and why-questions, one always asks why some fact obtained rather than another member of a contrast class of facts. If one takes the explanandum and contrast class to be the tower having a particular height as opposed to a different height—as Van Fraassen seems to—
one is immediately trapped into a causal explanation. However, if one instead considers the explanandum and contrast class \{the tower has height h and that is significant, the tower has height h and that is not significant\}, the correct explanation does invoke the later actual shadow length.\textsuperscript{123}

Notice that Van Fraassen was right to emphasize the importance of contextual factors determining what sorts of relations explanans can bear to explananda—sometimes later facts are admissible, whereas sometimes they are not. And if small variations in context can change the constraints on the \textit{explains} relation, that suggests that a full accounting of explanation might require abstracting away from particular constraints on structure or content. Pragmatic accounts, which eschew characterizing explanation in terms of particular structures or contents and instead focus on the role explanations play in particular contexts, are ideally situated to account for the surprisingly dynamic nature of explaining.

The account I advocate elsewhere (Chapter 3), that explanations are just those things that could, under appropriate circumstances, cause improved understanding of the explanandum, handles ARCHDUKE quite well; if I did not already have access to such information, I gain improved understanding of the importance of the assassination (and perhaps even the event of the assassination itself) as a result of learning that it sparked World War I. Such an account of explanation in terms of understanding has the potential to do more than just correctly classify explanations of the assassination’s importance—it

\textsuperscript{123} One could attempt to maintain Van Fraassen’s contrast class but work the significance of the shadow height into the relevance relation. This route would face trouble, however, as the fact that a particular height is significant would not favor the tower’s actually being that height.
can also show why such explanations do not fit the usual pattern. As much of our explanatory practice is geared to manipulation and control, the understanding we seek is typically caused by causal explanations in most contexts. However, there are other contexts, such as in a history class, where a different sort of explanation is merited. A constraint favoring causes as explanans, adopted without qualm in other domains of inquiry, is thus problematic if one attempts to generalize it to a constraint on explanation in general.

That there are situations where effects/later states can explain causes/earlier states is of no help to the proponent of D-N or unificationism, as neither theory is sensitive to the triggers for favoring one sort of explanation over another. Moreover, one might suspect that the failure of the rule that only earlier states explain later states to be fully general might pose a problem for the very causal accounts that were designed to always avoid asymmetry. In the next section, I argue that causal accounts avoid problems of asymmetry only by imposing structural constraints on explanations that rule out cases like the explanation of the assassination’s importance.

§4 Causal Accounts and Explanatory Asymmetries

Causal accounts of explanation state that one explains an event or state-of-affairs by citing at least one cause. (Such theories are often at pains to non-trivially account for which causes are explanatorily relevant, but for present purposes we can grant that every causal account is accompanied by a story that successfully enables us to differentiate
explanatory from non-explanatory causes.) There are other potential worries one might have regarding causal accounts, but in this section I am concerned with whether the machinery they employ to block standard counterexamples such as Pseudo Eclipse will be too blunt to allow in legitimate explanations such as ARCHDUKE. I look at James Woodward’s causal account, and argue that it cannot countenance ARCHDUKE as a legitimate explanation without making unacceptable sacrifices elsewhere in the theory.

On Woodward’s account, the reason to generally favor earlier states over later states is that when one intervenes on earlier states there will be a change in later states (or at least there would be if nothing interferes), but not conversely. The problem is identifying what does the work in avoiding what are known as backtracking counterfactuals. One would think that, counterfactually, the closest possible world where the moon is somewhere different on January 1st, 2013 is one where it was already somewhere slightly different in July of 2012—given that the world is law governed, the closest world where the moon was where it actually was in July but somewhere different in January would be one where a miracle caused it to deviate from its path. However, one would think that a world with a moon-moving miracle would be farther away than a world where small differences in earlier conditions caused the moon to be somewhere else in July. Thus, if causation is mere counterfactual dependence as determined by closeness of possible worlds, the later position causes the earlier one in exactly the same way that the earlier one causes the later one. Woodward blocks this inference by specifying that interventions are themselves part of the causal nexus, and thus propagate only forward in time. If interventions could affect events that took place before the
intervention, one would think the counterfactual wherein one intervenes to move the moon’s later position would be one where one also moves the moon earlier. Interventions only occur right before potential causes, and affect things only after. There is good empirical (and probably metaphysical) reason to suppose that the world really does work this way—one never sees an intervention occur after an event. Woodward’s causal account thus succeeds in giving a reason to accept Real-Eclipse as an explanation but reject Pseudo-Eclipse.

However, those very same constraints prevent Woodward from being able to countenance ARCHDUKE as a legitimate explanation. Since it is assumed that effects of interventions only propagate forward in time, I cannot intervene on the later war to cause a difference in importance for the earlier assassination. If Woodward were to loosen the restriction that interventions propagate only forward in time, then his account would also predict that the moon’s later position explains (and, worse still, causes) its earlier position. Thus, if ARCHDUKE is a legitimate explanation it serves a role with respect to Woodwardian causal accounts that traditional examples of asymmetry served for accounts such as D-N. Both groups fail because they do not have the resources to track the way different interests and contexts affect what we would count as a good explanation.
Chapter 8: Taking FE Out for a Spin

Suppose one accepts everything in this essay as the truth about explanation and understanding. There is still a huge amount of work to be done in the area. In this chapter I do two things: First, I attempt to fulfill a promissory note from the first chapter to the effect that the account of explanation here expounded could shed light on a debate of philosophical interest. Second, I look at to what further research questions my project most naturally leads.

§1 Some Surprising Results Regarding Theism

In the introduction, I claimed that having an account of explanation might help us settle questions of whether a particular posit is explanatory; specifically, I claimed that we could perhaps answer the question of whether theistic hypotheses would, if true, explain facts about the universe or merely re-describe them. Perhaps surprisingly, the answer depends on which theistic hypothesis is in question. According to FE, the existence of a deity would potentially be explanatory if it could be appealed to in an explaining act that could cause understanding in some person. Understanding would be caused in some person by an explanation if it enabled him to better manipulate his world or make better inferences about it.
If true then, the claim that there is a person-like deity who created the world and has periodic (or continuous) interactions with humanity would be explanatory. I would be able to function better in the world by knowing what sorts of actions are likely to incur the wrath of a vengeful deity, and make much better inferences about its cosmology. While the atheist could still very well maintain that there is no omnibenevolent, omniscient, and omnipotent being interacting with the world, it would on the present account be a mistake to preemptively disregard the hypothesis that there is such a being as necessarily unexplanatory.

However, the more abstract one’s concept of a god becomes, the less explanatory it is. Deism, it turns out, offers little to no explanation. Learning that the universe is the result of the will of some being, but that said being no longer interacts with it, does not in any obvious way lead to more successful interactions with the world, and, outside of abnormal contexts where inferences about the moment of creation are oddly salient, does not enable one to make any successful inferences either. Branches of negative theology, which claim explicitly that one should refrain from making any inferences pertaining to the creator of the universe, fare even worse (from an explanatory perspective). This result is perhaps somewhat surprising—the apparently more cautious, and arguably scientific conceptions of an abstract deity prove less explanatory than some arguably less sophisticated versions of theism that posit that the universe was created by an understandable entity not too different from you and me. This result could be read as a new application for Wittgenstein’s famous observation that “a nothing would serve just as well as a something about which nothing could be said.” (1953/2009, §304)
§2 What now?

Having argued for FE and applied it to at least one case of active interest, the next obvious question is where further research could be most fruitful.

One could always dig downward into the more fundamental notions invoked by URM and FE. These include representation, effective inference, and contextual relevance.

Representation in particular merits more careful examination. Despite the prevalence of representation-talk in the cognitive sciences, it has recently been called into question in philosophy (for extended critiques of “representationalism”, see Mathews (2007) and Price (2011)). To the extent that such anti-representationalists’ claims conflict with the use to which I have put representations in this dissertation, the present essay presents a potentially unseen cost to their work.\textsuperscript{124} However, it is not as obvious as it might seem that the sorts of representations to which I appeal need be anathema to anti-representationalists. Matthews, for example, distinguishes between “representations” and “Representations”; it is only the latter (capitalized) notion that he thinks is defective. Matthews’s target “Representations” are only those that satisfy a specific set of characterizing tenets (2007, p. 20-21). The notion of representation invoked in URM is, however, relatively thin. While the sorts of representations invoked no doubt need satisfy some of Matthews’s tenets—for example, they must be explicit and computationally

\textsuperscript{124} Even the opponents of “representations” acknowledge that the default view is that the success of cognitive science makes the legitimacy of representation-talk the default position. Matthews, for example, refers to the view he attacks as near-canonical (2007, p. 19) and is at pains to try to dispute its apparent support from cognitive science’s success (2007, Chapter 3).
discrete (tenet iii)—it is not at all obvious that they need satisfy others. There is nothing in the development of URM that commits to the representations in question being in “quasi-linguistic form” (Matthews 2007, p. 21) or to the claim that all thinking need to be inferential symbol manipulation on such representations. One interesting project would involve going through Matthews’s (and others’) arguments against representations one-by-one and seeing which, if any, are general enough to apply to the bare-bones notion employed by URM.

Along another dimension, one could likely fill a career working out the pragmatics of explanation and understanding attributions.

What I find more interesting though, are questions that move upward to more abstract domains. One pressing question is the relation between the sort of explanation envisioned here and the common practice of inference to the best explanation. Given what has been said about explanations, they seem necessarily tied (at least for us) to the sorts of things we understand and hence to the sorts of cognizers we are. Yet it would be surprising if the way we think happened to track the way the world is in some important sense. If something’s quality as an explanation depends on something which does not track facts about the external world, then quality of an explanation would be unlikely to be an indicator of the truth of the explanation. Therefore, inference to the best explanation seems to be on rough footing.\footnote{As Peter Litpon (2004, Chapter 8) notes, we could always make inference to the best explanation trivially sound by requiring that everything cited in an explanation be true. Doing so would, however, make inference to the best explanation wholly useless as a means for discovering information about the world.} Given our reliance on IBE, this is an unfortunate result. We should try to either justify IBE despite the reasons to be suspicious
of it, rehabilitate practice, or perhaps even find a new model to our practice that does not require IBE.

Another project is applying FE to special domains. Mathematical explanation in particular has attracted much attention in the literature, often as importantly different from explanations in science. Specialists in every field employ their own methods in discovery and explanation; it remains to be seen whether all fields can really be brought into the one general framework, or whether some fields will present insurmountable special challenges. Explanation and understanding of languages, while addressed briefly in Chapter 5, also merits a significantly more detailed investigation.
References


Koslowski, B., Marasia, J., Chelenza, M., & Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework *Cognitive Development*, (Special Issue on Scientific Reasoning)


191


Pitt, D. (2004). The phenomenology of cognition or what is it like to think that


Salmon, W. C. ([1971]). In Jeffrey R. C., Greeno J. G.(Eds.), *Statistical explanation & statistical relevance* by wesley C. salmon, with contributions by richard C. jeffrey and james G. greeno. [Pittsburgh]: University of Pittsburgh Press.


