The Shadows and Shallows of Explanation

ROBERT A. WILSON
Beckman Institute, University of Illinois, Urbana-Champaign

FRANK KEIL
Department of Psychology, Cornell University, U.S.A.

Abstract. We introduce two notions–the *shadows* and the *shallows* of explanation–in opening up explanation to broader, interdisciplinary investigation. The "shadows of explanation" refer to past philosophical efforts to provide either a conceptual analysis of explanation or in some other way to pinpoint the essence of explanation. The "shallows of explanation" refer to the phenomenon of having surprisingly limited everyday, individual cognitive abilities when it comes to explanation. Explanations are ubiquitous, but they typically are not accompanied by the depth that we might, *prima facie*, expect. We explain the existence of the shadows and shallows of explanation in terms of there being a theoretical abyss between explanation and richer, theoretical structures that are often attributed to people. We offer an account of the shallows, in particular, both in terms of shorn-down, internal, mental machinery, and in terms of an enriched, public symbolic environment, relative to the currently dominant ways of thinking about cognition and the world.

Key words: explanation, theories, concepts, division of cognitive labor, cognitive development

1. Introduction

Explanation is a river that flows through human life. Explanations are ubiquitous in human life, pervading the spectrum of our activities from the most simple and mundane ("Just going to the store to buy some milk") to the most sophisticated and unusual, such as the explanations one often finds in science and mathematics. Explanations are found across all cultures and historical time periods (Sperber et al., 1994). For example, in the history of medicine, lay people and medical experts alike have offered rich explanations for various diseases, explanations that were often wrong in many key respects, but which clearly were explanations (Lindberg 1992; Scarborough 1969; Magner 1992). Similarly, in cross cultural-research despite huge variations in the details concerning mechanisms underlying disease, the search for causal explanations represents a striking commonality amidst this diversity (Atran 1996; Maffi 1994).

Despite this pervasiveness and the centrality of explanation to human life, explanation remains one of the most under-explored topics in the cognitive sciences. In psychology in particular, explanation – how extensive and various our explanatory capacities are, what types of mechanisms underlie our those abilities, how these abilities develop – is a topic that has been mostly discussed only incidentally as researchers have investigated related phenomena, such as problem-solving, theory-formation, text comprehension, and expertise.

Minds and Machines **8:** 137–159, 1998.

The most developed discussions of explanation are to be found in the philosophy of science, from its twentieth century inception in positivist models of science – reaching its high-point in the extended studies of Hempel (1965) and Nagel (1961) – to the contemporary post positivist explorations of Salmon (1989) and Kitcher (1989). These discussions have covered a huge array of issues. To name three of the more heavily traversed: the role of laws in explanation; the relationships between causation, explanation, and prediction; and the interplay between the articulation of theories and the growth of explanatory power. As admirable as much of this work is, its focus imposes two mutually reinforcing limitations in understanding explanation from a cognitive science perspective.

First, the concentration in such discussions is on *scientific* explanation. Although the various accounts of explanation offered – from the classic covering-law model to Kitcher's (1989, 1993) explanation patterns – have made at times acute observations about explanation in non-scientific contexts, it seems fair to say that scientific explanation has served as a paradigm for such accounts of explanation. And not unreasonably, since the sciences have been rightly taken to be the source of our best explanations for a range of phenomena, and their further development our best bet for arriving at good explanations for phenomena currently beyond their ken. There is, of course, nothing wrong with such a focus when one's concern is with science. Problems arise, however, in extrapolating from explanation in the special institutional and cultural location of science to explanation in its natural and more common niche: everyday life. Scientific explanation usually works in a community of scholars who rely heavily on public forums for sharing and accessing explanatory insights; and a large set of explanatory standards, both implicit and often explicit in many sciences, would seem to guide the generation and growth of explanations in ways that do not occur in every day life.

Second, due both to peculiarities of the until-recently-dominant positivist approaches to the philosophy of science and to the (to many, strange) abstract proclivities of philosophers themselves, questions about the psychological and social realities that underpin explanatory phenomena have remained largely unasked (let alone answered) within these approaches. The well-known distinction between the *context of justification* and the *context of discovery*, initially drawn by Reichenbach (1938) but prevalent within philosophy of science more generally until the last 20 years, has come to set a boundary between the "logic of explanation" – the province of philosophers of science – and the (mere) psychology or sociology of explanation. While recent and various "naturalistic turns" in the philosophy of science reject this dichotomy (see Kitcher 1992), the sort of boundary that it circumscribed around philosophical explorations of science remains largely intact, even if the area enclosed is somewhat more encompassing. Distinctively philosophical concerns – a preoccupation with the form(s) that best represent scientific explanations, with questions about scientific explanation in general, and with the relations between explanation and causation, theoretical knowledge, and modeling

– persist. Again, this may be fine insofar as understanding scientific explanation goes, but it constitutes a limitation on the study of explanation in general.

The first two tasks of this paper are themselves perhaps "distinctively philosophical" in that we will try first to *identify* and then to *locate* explanation *vis-a-vis* a cluster of related notions. Our aim is to do so in a way that both maps this conceptual space and opens it up for more psychologically-oriented exploration.

2. Identifying Explanation

In its most general sense, an explanation of a given phenomenon is *an apparently successful attempt to increase the understanding of that phenomenon*. As simple as this claim is, it implies a number of features that explanations have, features that any account of explanation should satisfy:

(i) explanations are the product of something that we, individually or collectively, do; they are the result of our activity in the world, and so explanations will often have intentional and social presuppositions and consequences.

(ii) since they are attempts, explanations may fail, and they will fail just when they do not increase the understanding of the phenomenon they purport to explain. For them to be embraced as explanations, however, they must, at least transiently, appear to some group to increase understanding.

(iii) since their aim is to increase understanding, there will always be a psychological aspect to explanation.

(iv) since aiming to increase understanding is something that can be done with respect to oneself or with respect to an intended audience, explanations may, but need not, have a communicative and pragmatic aspect.

(v) since success in increasing understanding usually reflects an insight into how the world is, explanations help to create knowledge, to develop better theories and models of how things are, and to explain why they operate as they do.

(vi) since any phenomenon is inextricably bound up with many others, any given explanation will always have implications for phenomena other that which it initially attempts to explain. Thus, explanation increases understanding not just for its target, but inevitably for a larger sphere of often completely unanticipated affairs. (It is this spreading penumbra of insight that, we will argue below, helps account for an important asymmetry between prediction and explanation.)

On its own, this simple list of features of explanation hardly gets to the heart of the concept of explanation. Note, however, that the features themselves are those that have been emphasized by contrasting philosophical approaches to explanation. But each approach has tended to over-emphasize some of these features at the expense of ignoring or down-playing the others. This imbalance, in part, is why these approaches only ever give us the *shadows of explanation*. For example, classic positivists (e.g., Hempel, 1965; Nagel, 1961) have overstated (vi) in their view of the symmetry between prediction and explanation, and have done so while

virtually ignoring (i) – (iv). Pragmatic approaches to explanation (e.g., Bromberger 1966, van Fraasen 1977), by contrast, overstate the point made in (iv) and to some extent that in (ii) and so have little to say about either social and psychological dimensions to explanation (other than those invoked in communicative acts), or the growth of knowledge that explanations can contribute to. Realist approaches to explanation concentrate on what makes (v) true, and, apart from (vi) pretty much ignore the rest. Finally, social constructivists, insofar as they have any view of explanation, take (i) to represent *the* central feature of explanation, and have little to say of interest about the remainder of our list.

Explanation, then, might be seen as being characterized by an extensive array of features which, only as a coherent aggregate, can start to take us beyond the shadows. Yet most accounts have avoided such a "family resemblance" approach and tended to try to isolate one single defining criteria. In addition to this excessive focus on just one of many features, there has traditionally been an assumption that all explanations are essentially of the same type, regardless of the *explanandum*. This assumption also seems wrong to us. The structure and very nature of explanations may interact heavily with what sort of thing is being explained. The kinds of causal interactions central to say, biology, may require a very different sort of interpretative knowledge than those central to, say, physics. The canonical explanations in such domains as history, mathematics, mechanics, quantum physics, folk psychology, biology, and economics may all have their own nature. Or perhaps somewhat broader explanatory types, such as statistical, teleological, and intentional intersect with different facets of these natural domains, sometimes being applied in different ways to the same phenomena to yield very different insights. Although we have questioned the assumption of a simple, "explanatory essence" by pointing to various scientific domains, it is perhaps most clearly false when we move to consider both common-sense *and* scientific explanations together. Your explanation of why you'll be late home for dinner and a mathematician's proof of a theorem share very little.

We think that a full characterization of explanation requires an integrated and coherent account incorporating all six of the features listed above. Yet we have said nothing so far about four aspects of explanation that would seem to many to be at the very heart of what explanation is:

(a) the role of laws and observation in explanation
(b) the relation between causation and explanation
(c) the place of measurement and mathematization in explanation
(d) the structure of explanation.

These are, of course, issues central to philosophical discussions of *scientific* explanation, but in trying to cast a wider net than those discussions, it is appropriate that our starting point be neutral with respect to specific commitments about any of these. For example, laws and quantification certainly do feature in at least many scientific explanations, but at least *prima facie* play little role in common sense explanations outside of science. Working with a conception of explanation that

makes a claim about one of these issues would be to assume that scientific explanations are a sort of paradigm of explanation, and we have already suggested that such a view is mistaken, at least in its general form. Understanding the nature of (a)–(d) for any kind of explanation will be important, but the specifics may vary greatly. Even if there are important structural properties to, say, explanation in both scientific physics and folk psychology, these may have little in common with one another.

3. Locating Explanation

Even though psychologists and other researchers in the cognitive and behavioral sciences have had relatively little to say about explanation, they have had a fair bit to say about some related concepts. Psychological studies of hypothesis formation and testing, of prediction and discovery, and of reasoning have their own well-worked paradigms. For example, the 2-4-6 paradigm for investigating how people generate and test hypotheses (Wason 1968), and both the Wason card task and the Kahneman-Tversky paradigms (Wason and Johnson-Laird 1972, Kahneman and Tversky 1973) for exploring human inference represent extensively utilized experimental paradigms in psychology that investigate notions related to that of explanation. More recently Brewer and colleagues have explored psychological dimensions to the theory-observation distinction (Brewer, this volume), and the (still) burgeoning literature on the child's theory of mind has created a forum in which the idea of nativist and developmentally-unfolding "theories" has been debated (Wellman 1990, Wellman and Gelman, in press). For example, it now appears that preschoolers have considerable difficulties identifying how they came across knowledge and understanding unless it led directly to a novel behavior (Esbensen et al. 1997; Taylor et al. 1994)). In one especially compelling condition four year old children were taught essentially the same information either as new facts or as new behaviors, such as the meaning of Japanese counting words (facts) vs. how to count in Japanese (behavior). They learned both very well, but had sharply contrasting intuitions about the origins of such knowledge. They claimed that they had prior knowledge of the information when presented as facts (often saying they had always known it) but were much more likely to see the behavioral version of the information as something that they had just learned.

Interestingly but not surprisingly, when psychologists have gestured at the broader issues into which these investigations feed, they have often appealed to the literature in the philosophy of science. For example, the idea of the "child as scientist" is familiar from Carey 1985 (that of the "scientist as child, a more recent contribution of Gopnik 1996), and here there has been a productive exchange of ideas between developmental psychologists and philosophers on conceptual change, theoretical knowledge, and the nature of concepts. It is far less clear, however, exactly where explanation fits into this discourse and how it is to be approached in an interdisciplinary manner.

We consider three central notions: prediction, understanding, and theories. Each has a clear psychological component, although we will suggest that theories are best thought of as psychological *only in a derivative sense*. We shall consider the idea that these notions form a progression of increasing sophistication and depth, with explanation falling between understanding and theories. Moreover, we can think of these as natural human competencies: that is, we all predict, understand, explain, and theorize as part of our everyday activities, and could not have anything approaching human experience without such competencies. Recognition of them as competencies helps highlight their action component, a theme that will be important in understanding what everyday explanations are all about. To consider the progression, we need to more fully characterize the sense in which prediction and understanding are weaker than explanation, and theory stronger than it. We start with prediction.

3.1. PREDICTION

We are all familiar with cases where we can predict that something will happen even though we are unable to explain *why* it will. For example, many of us can and do reliably predict that our cars will start when we turn the ignition switch, but few of us are able to explain in any real detail just why this is so. It is not, of course, that we would be without words were we asked "Why?"; rather, it is that the sort of explanation that typically ensues will be likely to do little more than indicate the engine and battery as being connected up to the ignition switch in some way. (With car engines being increasingly electronic and computational in detail, the paucity of our explanations here has become more striking in the last few years.) The same is true of most commonplace gadgets that we use everyday.

Correspondingly, there are many cases where prediction is completely barren of any insight or explanatory power; this is especially so for many natural phenomena. I may know that a red sky at night portends good weather the following day without knowing anything more. I may know that one high tide is followed by the other in roughly 13 hours without any understanding of why. Similarly, many arbitrary social conventions may have nothing more than a predictive component, such as predicting that the fork will be on the left of the plate at dinner. Even complex or unfamiliar artifacts can have a purely predictive component. I might know that my computer tends to crash when I print from a certain program but have absolutely no notion as to why. I might notice than an unfamiliar tool is always at one orientation at one time of day and a different orientation at a different time but again have no notion why.

It is also true, however, that for many of these artifacts, we do have a level of "functional" explanation that can provide some insight to those who are less familiar with the object. Thus, we might well be providing some insight to a new adolescent driver when we say that turning the key starts the car because turning it starts an electrical sequence that directs energy stored in the battery to a starting

motor that turns over the engine, and which also allows all the electrical parts of the car to become active and interact. Or one might explain pressing a print key on a computer to a preschooler as telling the computer to send all the words on the screen to the printer, and as telling the printer to put all those words on paper. Even though our understanding of detailed mechanisms may be minimal, we can often provide some explanatory insight at this relatively shallow functional level. Moreover, we can use such functional explanatory knowledge to make predictions, troubleshoot a faulty system or to help evaluate more detailed explanations about how it works. We see this aspect of explanation to be a reflection of our desiderata (i) and (iv) above.

For any phenomenon for which we have an explanation, there seem to be myriad predictions that we can and do make – many of which are confirmed. Suppose that Joe is able to explain to you, in some detail, just why your car starts when you turn the ignition switch. Then Joe will also be able to predict what will happen if various parts of the engine are damaged or removed, and if the conditions under which you try to start your car are varied (e.g., radical temperature shifts), as well as what will happen in a variety of perhaps apparently unrelated situations (e.g., what will happen when you fiddle with bits of your lawnmower). An explanatory ability in a particular case may buy you predictions that you could not otherwise have; after all, predictions typically come in clusters and so an ability to predict one thing typically carries with it an ability to predict many other things. But whereas explanatory ability seems to provide predictive ability, predictive ability doesn't buy you explanatory ability. It is in this sense that explanation is stronger than prediction. This is, for humans at least, a *psychological* point, though we suspect that it will be true for any predictive-explanatory device.

This intuitive point about the relative ease with which prediction can be generated seems confirmed by the existence of automated predictive devices in a range of fields, from predictions of the locations of aircraft in a control tower to predictions of the likelihood that a treatment will halt the progress of an otherwise fatal disease. Even powerful and impressive predictions can be generated simply from correlation and simple inductive strategies of extrapolation. Whether any of these devices can be said to explain the corresponding phenomena seems to us doubtful. The case becomes somewhat more intricate when patterns of covariation are considered (Cheng 1997; Glymour, this volume; Thagard, this volume), but here too we doubt that we have explanation.

Even though explanatory ability guarantees predictive ability in the sense specified above, this does not imply that explanatory ability always allows us to predict corresponding specific outcomes. Prediction of a specific future state of affairs from a present one may be practically impossible, while a full explanation after the fact might be quite easy. Informally, we are often in situations where we might say that we could never have anticipated that something would happen but that, after the fact, we see exactly how and why it came about. More formally, this effect is related to the difference between physical systems that can be modeled by simple

sets of equations and those that while fully determinate, cannot be so modeled. In this sense explanation is *easier* to come by than prediction, and it provides us with further reason to create some cognitive distance between prediction and explanation. What we have in mind here can perhaps best be conveyed through examples.

In folk psychology, we can often explain why a person did what she did after the fact, even though we could not predict ahead of time how she would act. This is not simply because we are not in a position to observe the person sufficiently, but because just how people will behave is influenced by a myriad of interacting variables, the values of which may depend on minute details of the situation. This is surely why so many of our conversations about the minds of others are attempts to make sense of a behavior after the fact rather than to predict it. Such "retrodictive" explanations are not just the province of psychology; they are seen throughout life. The crew of a commercial swordfishing boat may be completely unable to predict prices for the fish they caught when they return to port. Yet they may easily explain how and why the prices dropped when another large boat half way around the world just checked in to a port before them with a huge cargo of swordfish. Or, one may not have any ability to predict the sex of one's child, but might well be able, after the fact, to be able to explain how that particular sex was the result. In all of these cases, although anyone in a position to offer an explanation will also be able to make some corresponding prediction (minimally, about what would happen in certain, counterfactual situations), explanatory ability here serves a function independent of the generation of predictions.

There has been much attention in recent years to the properties of fully deterministic non-linear dynamic systems, where extremely minor differences in initial conditions can lead to dramatically different outcomes (e.g., Waldrop 1992). Since we cannot often know such initial conditions to the necessary levels of precision, we are pragmatically unable to make predictions for any future events critically dependent on those initial conditions. Part of our difficulty in being able to predict turns on the time frame of the prediction. I may be able to predict the weather in three hours at levels far above chance but may have not predictive power for three weeks in advance even if, at that time, I can explain how the weather got that way. But often the time frame of most interest is just the one we cannot predict in advance but can explain after the fact. One may be intensely concerned about how a volatile relative, Uncle Jack, will react at an upcoming wedding, but be unsure as to whether Uncle Jack will be extremely enthusiastic about the wedding or profoundly offended. Which behavior emerges may depend on the precise wording of an offhand remark made by another family member, a remark that might be taken as an insult or a complement regardless of the speaker's intentions. Uncle Jack's 'threshold of defensiveness' may have a hair trigger sensitivity that makes it impossible to know in advance which way a remark will be taken; yet the ensuing behavior of either type may make perfect sense to those who understand the underlying dynamics of Uncle Jack's personality. In addition, both the explanation and

the dynamics may be considered absolutely essential to any real understanding of Uncle Jack's behavior. Thus, explanation after the fact may be most powerful just where prediction is weakest and least effective.

Discussions of explanation seem to often underestimate the extent to which most real world events are non-linear dynamic systems in which the grains of analyses at which we do the most explaining often contain almost no specific predictions. Yet, if such explanation-without-prediction cases are so common, one wonders why we even engage in explanations in such cases. One possibility is that explanations help sharpen the ability to perceive and respond to events in the future. Consider, for example, what happens if one provides someone with glasses that greatly improve their vision. It would be odd to say that the glasses help them predict events better; but they might certainly help them pick up information more accurately and powerfully. The lenses greatly improve the quality of information gathered in real time, thereby allowing richer and more powerful interpretations of experience and the ability to interact with aspects of the world more effectively. Like lenses, explanations may often serve to sharpen our perceptions of events, to be able to see more clearly what is going on in real time without necessarily being able to make better predictions for more than a moment or two in the future. Explanations serve to buttress our overall conceptual frameworks for interpreting and making sense of the world around us.

As a final example, consider those cases where a fanatical sports fan takes a friend unfamiliar with the sport to an event, such as a hockey game. The fanatic will certainly make some predictions his friend will not, but much of the time he can not better predict the play that will happen in the next few minutes than his friend. Yet he is vastly better at understanding what has happened and what it meant. So also for the players in the event, who might know with great skill what a particular situation "means" and how to best to respond to it, without knowing what the situation will be in a few minutes.

We are not claiming that increasing explanatory insight carries with it no predictive gains; but we do want to suggest that such gains may often be in a very narrow time window while the largest explanatory insights may be over a much larger time frame. The value of that insight over the larger time frame lies less in prediction and more in the ability to "see" and remember the dynamics of future events more accurately. Explanations thus can provide better lenses on the causal structure of the world.

In short, explanation seems much conceptually richer than prediction and typically entails certain kinds of predictions. But explanation may be functionally important even when it does little to predict the phenomena it explains the best.

3.2. UNDERSTANDING

Subject to the caveat above, understanding entails prediction but also includes accompanying knowledge through inference and memory, and some sense of how

and why things happen, even if this remains largely implicit and inarticulate. It is also less articulate and communicative than explanation, since one can understand something without being able to explain it to another; the reverse, however, seems impossible.

We want to suggest that understanding, which may otherwise be largely implicit, must be made explicit for either the communication to others or the reflection by oneself that typifies explanation. The existence of an explanatory sense or hunch that is more than understanding but not so explicit as to be propositional supports this suggestion. Thus, one might be able to choose appropriately between competing explanations for reasons that aren't obvious even to oneself because one just seems to "fit" better.

Our conception, then, is of understanding as a cognitive state that remains largely implicit but which goes beyond merely being able to correlate variables. We think that one chief source for understanding, so conceived, is that as agents in the world, cognizers are often in a position to have some sense of how and why things happen through knowledge of their own actions. Notoriously, we often have *know-how*, or procedural knowledge, largely implicit and non-reflective understanding that goes beyond simply being able to predict the co-occurrence of two or more variables (Zola-Morgan and Squire 1993). Any of the basic activities of daily life carries with it a sense of the place of that activity within the surrounding causal nexus that is not merely predictive. For example, in walking through a neighborhood park you perceive many things and events familiar to you – dogs being walked, children on swings, the swaying trees – each of which you have some understanding of, even though you may not have articulated any thoughts about them. Moving from understanding to explanation involves that further articulation. Note also that the Taylor et al. studies on the development of awareness of knowledge referred to earlier in this section suggest that knowledge that leads to new actions is identified earlier in development as something that one has learned and can be evaluated as such more easily.

Understanding therefore might often have the flavor of procedural knowledge that occurs outside the sphere of conscious thought. One can understand how to tie one's shoes without being able to explain it to another. Indeed, this is often the problem when world class athletes are hired as coaches. They might understand perfectly *how* to execute certain complex moves but be unable to explain them at all to novices.

3.3. THEORY

Finally, we turn to the relation between explanation and theory. The idea that individuals either construct theories, or even maturationally develop them, and that this is what enables them to engage in explanation of the world around them, underlies much contemporary research in cognitive development and work on the psychology of science. The standard experimental paradigm used to explore this

idea provides the participant with a task, the response to which is best explained by that individual's possession of a theory regarding the domain into which the phenomenon investigated by the task falls. For example, when infants preferentially look at an object that violates simple constraints on bounded physical objects, such as not being interpenetratable, it is common to attribute to them a 'theory' of the mechanics of objects (e.g., Spelke 1994). Or when 4-year-olds attribute distinct sets of causal relations to only living kinds, one might claim they have a naive theory of biology (Keil 1995). Our view is that this attribution of theories to individuals has been too free and easy, in part because of an overlooked feature of everyday explanations: how shallow they are.

We can begin elaborating on what we mean by the shallows of explanation by extending the metaphor we began the paper with: that although explanation flows through nearly all forms of human life, it typically does not run very deep. More concretely, explanations typically stop or bottom out surprisingly early on. To return to one of our earlier examples, although almost everyone who owns a car can give some sort of explanation as to why their car starts (or doesn't) when the key is placed in the ignition and turned, few of us are able to respond with any depth to even the next few follow-up "why" or "how" questions. And the shallowness in this case is the norm: we rarely have ready access to explanations of any depth for all sorts of phenomena for which we are able to offer some sort of explanation. Indeed, we often carry with us an illusion of depth until we are faced with the actual task of explanation. Thus, people frequently seem to think they have vivid, fully mechanistic models of how something is or how it got the way it did, but when forced to state explicitly that mechanism as an explanation, their own intuitions of explanatory competence are shattered. For example, in current research in the second author's laboratory, college students are asked whether they know how various familiar devices work, such as toilets, contact lenses and bicycle derailleurs. Many assert that they have a complete and fully worked out understanding such that they could explain all the necessary steps in any process involving the object. Yet, when asked for such explanations, a large percentage of these participants will show striking inabilities to put together a coherent explanation, missing not just a few arbitrary details, but critical causal mechanisrns. Until they attempt such explanations, they are often under the illusion that they have a complete, "clockworks," vivid understanding.

We can distinguish two different kinds of shallows, those within a level and those across levels, where we can think of levels either as those in the classic reductionist hierarchy (physics, chemistry, biology, psychology, sociology), or, perhaps more perspicuously, in terms of Simon's (1969) nearly decomposable systems. The two different kinds of shallows represent distinct ways in which explanation may be circumscribed or limited. To take an example of the shallows of explanation across levels, while one might have a perfectly detailed mechanistic account of how an electric blender works within a level – in terms of gears, blades, motor torques and the cutting action of the blades on solids at different velocities – one's knowledge

across levels – knowledge of how electromagnetism makes a motor work, or how the chemical bonds of solids change when they are mechanically sheared – may be extremely limited, and hence explanations here are shallow. Conversely, one can also have huge explanatory holes at the primary level of analysis, as with the car starting example, or when most people try to explain the motions of gyroscopes (Proffitt and Gilden 1989), even if one is able to provide detailed explanations across higher and lower levels for particular components of the overall system.

On our account, theories have much more depth than explanations, but this greater depth does not mean that theories must explain everything relevant to a phenomenon. The normal usage of the term "theory" has never required a "theory of everything". For example, a theory of how the agent for mad cow disease works does not require a regress down to particle physics to be a legitimate theory. Theories have more scope and systematicity than explanations, but can reach a kind of natural boundary when they tie together a wide range of phenomena in one coherent account that links both observables and unobservables. This boundary relieves one of the requirement that theories be exhaustively deep (Wilson 1994, 1995, Chapter 8).

Just as the shadows of explanation are owed to explanation's ubiquity, the shallows of explanation are a consequence of the frequent occurrence of explanations in the absence (or minimal presence) of theory. It is in this sense – precisely that which we invoked in arguing that explanation is stronger than (mere) prediction – that explanation is weaker than theory.

Let us make our argument here more explicit:

1. Explanation is typically shallow.
2. Theories allow one to offer explanations with more depth than the shallows of explanation suggest that we typically provide.
   Therefore,
3. Having a theory about X entails being able to explain X
   but
4. Being able to explain X does not entail having a theory about X
   Thus,
5. Explanatory ability is weaker than theoretical ability.

The first premise is, we claim, itself a phenomenon in need of explanation. The second premise is a fact about theories expressed in light of the first premise. The inference from these premises to our initial conclusions follows provided that we can eliminate hypotheses that claim that the shallows of explanation are caused by something other than what we will call the *theoretical abyss*.

4. The Shallows and the Theoretical Abyss

We have suggested that, often enough, a theoretical abyss exists between our ability to provide limited explanations that suffice for the purposes at hand and the possession of corresponding, detailed theoretical knowledge that would allow

us to provide more satisfying, richer explanations. There would, of course, be no theoretical abyss if the shallows of explanation were owed to something other than the absence of such theoretical knowledge, and so here we shall consider why alternatives to the theoretical abyss are not all that plausible. In particular, we argue that a range of alternatives that appeal to social aspects of explanation and to general processing limitations should be rejected. So we will be arguing that the following sorts of hypotheses are false:

H_1. The shallows of explanation are simply a function of contextual or social features of the practice of explanation
and

H_2. The shallows of explanation stem from a limitation in our abilities but are a consequence of general processing and access abilities, not the absence of theoretical knowledge.

H_1 and its variants seek to identify the shallows of explanation as a sort of shortcoming of social performance or the pragmatics of communications (e.g., the maxim of not giving too much information in discourse; see Grice 1989). Personality traits (e.g., shyness) and the level of social comfort will certainly account for some cases of why explanations are shallow. But such hypotheses are implausible in the case of explanation more generally, since the shallows of explanation are a feature of both communicative and reflective explanation, where only the former need involve social performance at all. In the reflective case, we may often think about a phenomenon, decide we know how it works and then file that "explanation" away without communicating it to others. Precisely because we don't explain it to others, we may further entrench the illusion of explanatory depth. Perhaps we even confuse a sense of understanding with having a true explanation, or perhaps we have explanatory fragments that seem so clear that we falsely assume we know all the links between those fragments as well.

Like H_1, H_2 and its variants also views the shallows of explanation as a performance limitation, one that is due to memory and processing limitations. But they are implausible because the shallows of explanation manifests themselves not only in contrived experimental situations or cases where bottle-necks are imposed through task demands; they are pervasive in explanation "in the wild". To insist that we have the theoretical knowledge that would allow us to overcome the shallows of explanation but don't draw on it for, broadly speaking, reasons of cognitive architecture would be plausible were there circumstances in which we *didn't* fall into the shallows of explanation. But, so far as we can tell, the *only* way of avoiding the shallows is to *learn a theory*, i.e., to acquire precisely the web of knowledge that, we claim, is typically missing. In addition, there is every reason to believe that people can and do know causal propositional structures vastly more complicated than those required to escape the shallows of explanation. After all, for millennia humans have shown fabulous abilities to accurately remember lengthy narratives with complex internal causal structures, logical arguments, and entailments, and carefully laid out presuppositions and assumptions that lead to predictions. Perhaps

the overall cognitive demands of explanation are radically different from those of learning a narrative, but we see no signs of such a difference.

Of course, we have only considered two of the more obvious alternatives to the existence of a theoretical abyss as an explanation for the shallows of explanation, and so would be rightly accused of posing a false dilemma (trilemma, actually) if we were to rest our case here. But we think that the theoretical abyss also comports rather well with some broader features of explanation and the sort of division of cognitive labor it invokes, and it is to these that we now turn.

5. The Division of Cognitive Labor

We rely on knowledge in others extensively in our explanatory endeavors, and we rely on the assumption of knowledge in others to give us a sense of explanatory insight. This division of cognitive labor is a critical, prominent part of everyday explanation, one whose typical omission in discussions of explanation is, we think, a consequence of overlooking the shadows and shallows of explanation.

What do we mean by a division of cognitive labor? Putnam (1975) introduced the idea of a linguistic division of labor in his argument that "'meanings' just ain't in the head", and we base our conception on his. Putnam's idea was that while everyday users of natural language are able to apply the terms of that language by knowing what he called *stereotypes* of those terms, it is only "experts" who know the real essences of the referents of those terms. There is thus a sort of division of linguistic labor between the folk and various sets of experts, whereby the folk make do with relatively superficial referential knowledge but are still able to talk about the kinds of things there are because of the knowledge that experts have of the "essences" of those kinds of things. To take Putnam's most famous example, while everyday folk know that water is a clear, drinkable liquid found in lakes and that falls from the sky when it rains, it is only experts who know the underlying molecular essence of water, i.e., what water *really* is.

We propose that there is a similar division of cognitive labor that underwrites explanatory knowledge. That is, everyday folk know enough about the "nominal essences" of the things that they interact with on a regular basis in order to be able to offer relatively shallow explanations for their behavior. But there are also experts who have either the within-level or across-levels knowledge that the folk typically lack, and who are in a position to offer explanations with more depth. Although *we* are faced, as individuals, with the theoretical abyss as the norm, the theoretical knowledge that we lack exists somewhere, just not in our heads. This is to say that explanation and the theories that underwrite their depth, are *wide*, i.e., they do not supervene on an individual's intrinsic, physical properties.

The extent to which theories and explanation are wide is even more striking than that of meanings (Putnam 1975) and concepts (Millikan, in press). Explanations are intimately linked to the structure of the world they try to account for and to the broader community of knowledge. Explanations, far more than for meanings

or concepts, are expected to work in that they should enable us to interact more proficiently with some aspect of the world. For that reason, they must strike a resonance with some aspect of the real world. We assume the following:

A. The structure of the world is organized into clusters with their own distinctive levels and kinds of patternings, causal and otherwise.
B. To be able to get much traction in thinking about those regularities, theories and explanations must be specifically tailored to the structures in each of these clusters of domains.
C. This specialization means that theories and explanations will be different, not just in what they refer to, but in their structure and form as a consequence of what they are trying to explain.

Just as the different sense organs have evolved very different sorts of structures for processing such different patterns as light and sound, theories of biology and physics are different from the bottom up. To understand them and how they work one must see them as linking a person to the world, not just to an internal mental representation. To handle a variegated perceptual world, we have evolved distinct perceptual modules; to handle a complicated theoretical world, we have enacted a form of distributed cognition in the form of the division of cognitive labor.

How could this division of cognitive labor work? It could not be that we simply have labels for various kinds of experts, such as physicists, chemists, biologists and doctors, for just knowing those labels would do no work. We must have some sort of insightful sense of what goes on in those areas of expertise, that is, of how mental constructs in those experts relate to the things they know so much about. The shallows notion may be the key here as it gives us an ability to know, in a superficial way, what explanations are like in a domain without really knowing much at all in the way of detail. This is far different from the normal sense of distributed cognition (e.g., Hutchins 1995), but it may be the central one to understanding how explanation works in broader social contexts.

We see two complementary ways in which the division of cognitive labor could work: through schematic modes of construal, and through public forms of representation, somewhat like “blackboards” that Simon suggests (this issue). The modes of construal allow people to have some sense of what experts know in a domain without knowing the details. The deep reliance of explanatory practices on public forms of representation – from writing systems, to iconistic symbols, to video-displays – implies that what constitutes or realizes an explanation literally extends beyond the head of the individual (Wilson 1998, typescript). We spell out how such notions might work in Sections 6 and 7 below; both reflect the sense in which explanations are not “in the head.”

The shallows notion also suggests a different view of what concepts are and how they fit into explanations. There has been great attention of late to the “concepts in theories” view of concepts (Carey 1985; Gopnik and Wellman 1994; Murphy and Medin 1985). But so much of that discussion has seemed to assume a view of theories as those kinds of concrete detailed models of reality that we have just

argued are so uncommon and which are usually impractical and not useful. Instead, if we think of concepts as embedded in "modes of construal", we start to see that their role in explanations can fit very nicely with the shallows idea.

We have argued that the shallows of explanation are themselves to be explained by the distinction between explanation and theory. This distinction comports nicely with our rejection of a concepts-in-theories view that requires explanations to be explicit, propositional entities (whether spoken, written, or thought) that contain concepts. Precisely what concepts are is a thorny issue on which we do not propose to take a stand here. But we do want to address the question of the extent to which concepts, explanations, and theories all involve irreducible causal mechanistic aspects, and to advance a position that is compatible with our views on the shadows and shallows of explanation.

6. Irreducible Causal Explanation and the Shallows

It is striking that the notion of a cause is invoked in almost all everyday explanations, whether directly or via the notions of causal structure, relations, and powers. From explaining why the water boils in the kettle when you turn the stove on, to explaining how trees grow, to explaining what people who join a health club seek: we find causation almost everywhere in explanation. Exceptions include purely mathematical explanations, explanations that appeal solely to logical features of a situation (such as inconsistency), and discussions of some legal and social conventions. We shall concern ourselves here solely with causal explanations, noting that we construe this notion broadly to encompass most explanations that we encounter in both everyday life and science.

How is the pervasiveness of *causal* explanation compatible with the shallows of explanation? After all, if causes are either explicitly or implicitly invoked in everyday explanations, then explanations must have some sort of depth to them, since causes are, often enough, underlying entities, and are, often enough, not themselves observed. In short, to put this more pointedly, causes are often *theoretical* entities, and their postulation thus presupposes the existence of theories of some sort, however impoverished. Given that, the prevalence of causal explanation seems incompatible with the theoretical abyss that we have posited in our account of explanation.

Our view is that while causes *are* invoked in explanation all the time, it is *how* they are invoked and *who* gets to invoke them that is the key to resolving this puzzle. Moreover, this appeal to the ways in which causes are drawn on in explanation, and by whom, not only provides an understanding of the way in which the ubiquity of causal explanation is compatible with the shallows of explanation, but also points to another partial cause of the shallows. The concept of cause that is ordinarily appealed to in explanation is not much more than that of "something that brings about, in some way, the phenomena that we seek to explain". It is truly a "we know not what", to use Locke's characterization of substance in the *Essay*

Concerning Human Understanding. And those who do know about the nature of the relevant causes are often not the ones offering the causal explanation. The "how" and the "who" here correspond roughly and respectively to the two ways in which the division of cognitive labor function: via sketchy modes of construal and the extended minds that result from cognizing through shared forms of representation.

This feature of our appeal to causes helps to explain both the shadows and shallows of explanation. We take the notion of causation itself to be a primitive notion, one that has its own shadows in both reductive accounts of causation (e.g., Humean accounts) and nonreductive accounts that consider causation to be richer but still analyzable in terms of prior notions (e.g., time, powers, properties). Causation, like explanation, is ubiquitous and best understood as a cluster notion; hence philosophical reconstructions of the concept are doomed to a shadowy existence. We can, perhaps, even see the shadows of at least causal explanation as inherited from those of causation itself. But we also appeal to causes even when, in is a very real sense, we don't know what these are; we seem almost perceptually built to infer the presence of causes, even if the theoretical understanding necessary to understand the nature of those causes lags far behind. The Humean cues of contiguity and constant conjunction are often sufficient for us to suspect a cause, irrespective of whether we have any conception of the type of mechanism involved or the underlying character of that cause. Given the shallows of causation, it is no wonder that our causal explanations are themselves shallow!

7. Shadows, Shallows, and Explanatory Success

The problem we have gestured at in the previous section has a general form that we would like to address more fully in this section. We argued in Section 2 that the shadows of explanation are a reflection of the inherent complexity of explanation and the difficulties of understanding it from any one perspective. Explanations may all share a function of helping their users at least think they understand a phenomena better, but beyond that very broad functional description, there are a huge array of phenomenological variations: flashes of insights, slow creeping realizations, picking up on a useful analogy, narrowing down alternatives, and so on. The psychological experience of explanation thus itself has a shadowy nature (for an alternative view, see Gopnik, this volume).

This shadowy nature, however, may not reflect a loose family resemblance concept as much as a rich implicit structure to explanations that is not easily translatable into explicit formulations. We think this large implicit structure is also linked to the issue of the "shallows of explanation" discussed in Sections 3 and 4. The shallows represent the surprising extent to which explanations do not explain many phenomena in very many steps.

Why, then, do explanations work for us if they are so shallow, so devoid of detailed mechanisms most of the time? Is there anything systematic about their structure that enables them to work? We argue that there may be several patterns to

explanatory knowledge that give a framework that allows us pick between classes of explanations without knowing much at all about specific mechanisms. And, given that the potential depth for explanations of many natural phenomena is so vast as to approach infinity, it may be a highly adaptive way for humans to gain and use explanatory insight. Here, then, is our general problem: how do we get a handle on the patternings that exist in the world for them to be of any use to us while not having clear notions of mechanism?

Here are some ways in which we might have an explanatory sense but stay firmly in the shallow end of the explanatory pool:

a. We can have senses of *explanatory centrality* of particular properties in particular domains. By this we mean a sense that some kinds of properties are particularly important in some domains. Color of an object, for example, figures more centrally in explanations involving most natural kinds than it does in explanations involving most artifacts (Keil, et al.). Size of an object may tend to impact more on artifacts as it can disrupt function more. These notions would be relatively useless if they had no generality and differed for every small level category. Instead, however, it seems that there are strikingly common patterns at a very high level. Thus, all animals tend to have roughly the same sorts of properties as explanatorily central and these will be very different from those for artifacts or non-living natural kinds (Keil and Smith 1996; ibid).

   There is a great deal we need to understand more fully here with respect to different sense of centrality (Sloman, et. al. in press). Consider a property's causal potency, i.e., the extent to which changing a property causally impacts on other properties of a kind, destabilizes that kind's integrity etc. This may be one of the most powerful and intuitive senses of centrality. But there is also the sense of non-causal centrality in terms of a key construct in mathematics or social conventions. One question asks if causal potency works at a more general level than the other forms of potency and centrality.

b. Notions of causal powers have been prominent in recent years in the philosophy of science, especially in the philosophy of biology (e.g., talk of gene action, and debates over the units of selection) and in the philosophy of psychology (e.g., the individualism debate, mental causation). These notions too can give an explanatory sense without yielding precise mechanisms. Troubling here is that the idea of an object's causal powers is used in a loose and often ambiguous way, as one of us has argued previously (Wilson 1995, Chapters 2, 5). More pressing in the current context is that we don't have a really clear idea of what notions of causal powers amount to at the psychological level. They seem stronger than notions of causal potency because they can be so specific to particular properties and very low level categories. It might seem that "causal powers" is another expression for a property or object, but the real sense seems more one of an interaction between a kind of thing and the world in which it is situated. Thus, we can understand and explain something

in terms of its causal powers, which means not just listing its properties as set of things attached to it, but, rather, listing its dispositions to behave in certain ways in certain situations. A hammer has the causal powers to pound in nails and remove them, and thinking of it as a decontextualized 'pounder' seems to miss the point. One has to think of its causal powers in terms of the sorts of things they act upon. Causal powers then seem often to be conceived of relationally, rather than as intrinsic properties that can be simply abstracted from the contexts in which they are instantiated.

So we might well have strong constraints on what count as appropriate explanations in a domain that come from causal powers notions without having specific mechanisms in mind and thereby still remaining in the explanatory shallows. I know that gold has a wide array of causal powers that are distinctive to it and expect that any explanation involving it must be in accord with those causal powers. But at the same time I may have little or no understanding of why it has those causal powers. Much the same may also be true for attributions of causal powers to people, animals, and artifacts.

c. We have a sense of explanation that is based on notions of kinds of agency and kinds of cause. Thus, I can think that certain kinds of agency and cause are much more likely to be central in one domain than another. Intentional agency is critical to understanding humans but not earthquakes. Teleological agency is more critical to biological kinds. Similarly, action-at-a-distance may be a kind of causality we expect to dominate in both psychological and gravitational interactions, but not mechanical ones. The explanatory value of such notions depends critically on how fine grained and reliable they might be, topics that are still hotly under debate; but again, even with rich detail, one could still not really have a clear sense of specific mechanism.

d. Related to kinds of agency and cause are notions about kinds of causal patternings. But these are importantly different and need to be understood as such. Independent of kind of agency or cause, might be patterns such as whether causal interactions proceed in serial chains or are massively parallel (the former perhaps being more common in simple artifacts), or whether many properties converge to support one or diverge from a common source to support many others. There are a large number of such patterns that one can identify and associate with particular domains, but again only as frameworks or guidelines.

Taken together, we can think of these four aspects of the shallows of explanation as yielding modes of construal that help us take an explanatory approach to a problem without really knowing all the details, and perhaps never doing so. These modes may be what drive not just most lay intuitions but those in science as well (Dunbar, 1994). Moreover, they may often be implicit in ways that make them a presupposed background in many scientific discussions.

8. The Shallows and Developing Explanation

That modes of construal and the shallows could work so well in helping us get an explanatory sense from others' expertise and know-how, and when to access that expertise in more detail can be seen from looking at how such notions emerge in all of us. In particular, it is beginning to appear that even before children have entered elementary school, they have a strong sense of how knowledge is clustered in the minds of experts, and it seems they must do so through navigating the shallows and using modes of construal. For example, preschoolers seem to at least have notions of causal potency/centrality, causal patternings, and kinds of agency, and almost surely of causal powers as well, although that notion has not been investigated systematically (Keil et. al., Wellman and Gelman). Causal potency is seen in preschoolers in their abilities to know that some sorts of properties are likely to be much more central in some domains than others. Again, color is understood as more likely to be central to natural kinds than to artifacts (ibid). This has been looked at primarily in terms of the extent to which counterfactual statements are seen as undermining a kind's integrity (e.g., a red-tire-looking-thing is a still a tire, but a red-seagull-looking-thing might well not be a seagull). Ongoing research is now asking how such notions of centrality would influence young children's preferences for some explanations over others.

There is also evidence that young children have senses of different causal patternings in various domains. Thus, they seem to know early on that action at a distance is a more reasonable kind of causal pattern for animates than inanimates (Leslie 1995); or that some patterns of causal homeostasis may fit better with natural kinds than with artifacts (Keil 1995). They also understand that the agency responsible for purposeful movement in plants is different from that in sentient beings. Thus, the sunflower follows the sun all day because of a very different kind of agency than that in the human sunbather.

Most recently, research in the second author's laboratory is showing that preschoolers have strong senses about how pieces of explanatory knowledge might be clustered in the minds of others – exactly the sort of understanding that would be central to a working division of cognitive labor. For example, a child might be told that Bill knows all about why two magnets, if turned the right way, stick together; and that John knows all about why a china lamp breaks into pieces if it falls off a table. The child is then asked who knows more about why television screens get all fuzzy sometimes during thunderstorms. Even preschoolers will cluster explanations about electricity and magnetism together to a greater extent than either of those explanation types with mechanics. There is no doubt that they are in nearly full ignorance of any specific mechanisms, yet they have some sense of how some explanations are more likely to be related in the minds of experts. In this example they may be keying into notions of invisible forces and action at a distance. Our general point, however, is that throughout much of development, and long before formal schooling, a set of framework explanatory schema are at work and seem to

be essential for further theory growth and conceptual change. The need for such structures so early in development may be yet another reason why the skeletal 'shallows' format is so psychologically important.

9. Conclusions

We have introduced two novel notions – the shadows and the shallows of explanation – in embarking on the larger project of opening up explanation to broader, interdisciplinary investigation. The "shadows of explanation" refer to those philosophical efforts to provide either a conceptual analysis of explanation or in some other way to pinpoint the essence of explanation. The "shallows of explanation" refer to the phenomenon of having surprisingly limited everyday, individual cognitive abilities when it comes to explanation. Explanations are, as we said at the outset, ubiquitous, but they typically are not accompanied by the depth that we might, *prima facie* expect.

We have attempted to explain the existence of the shadows and shallows of explanation in terms of a theoretical abyss between explanation and richer, theoretical structures that are often attributed to people, and thus suggested that the shadows and shallows of explanation are linked. In particular, if explanations are understood as largely implicit, skeletal notions about causal pattern – causal schemata, if you like – they will lead to both shadows and shallows effects. We see the shallows of explanation not only as compatible with humans' remarkable explanatory successes, including our grasp of causal explanation, but itself a reflection of the shadowy and shallow grasp we all have of causation. It further seems that this implicit skeletal format may be essential for two reasons.

First, it is the only way to cognitively handle the theoretical abyss; and second, it is perhaps the only format that could be mastered by the very young child. For other reasons having to do with how concepts emerge in development, the explanatory schema are critical early on, and the younger the child the more implausible any explicit fully detailed set of explicit propositions become. But beyond children, all of us find tremendous value in not having to master the full causal details in any domain. Instead, we get along much better by using our modes of construal and our social blackboard of signs and markers to access just the amount of depth we need on each occasion. Thus, we have offered an account of the shallows both in terms of shorn-down, internal mental machinery, and in terms of an enriched, public symbolic environment, relative to the currently dominant ways of thinking about cognition and the world.

To carry the shallows metaphor further, we know that we cannot dive infinitely deep or even stay in any depth for very long; but by using appropriate public charts and supporting frameworks, we can make occasional brief and directed dives, sometimes of surprising depth. So also, in focused and limited ways, we can go to extraordinary explanatory depths with the help of public charts and frameworks of

knowledge. But we could never possibly stay at such depths at all times across all domains.

10. Acknowledgement

Preparation of parts this paper and some of the studies described therein were supported by NIH grant ROl-HD23922 to F. Keil. Thanks to Leon Rozenblitt for helpful comments on an earlier draft of this paper.

References

Atran, S. (1996), 'From Folk Biology to Scientific Biology', in D. R. Olson and N. Torrance (eds.), *Handbook of Education and Human Development: New Models of Learning Teaching and Schooling*. Cambridge: Blackwell.

Brewer, W. C. Chinn and A. Samarapungavan, (1997), 'Explanation in Scientists and Children', *Minds and Machines*, (this volume)

Bromberger, S., (1966), 'Why-Questions', in R. Colodny (ed.) *Mind and Cosmos*. Pittsburgh: University of Pittsburgh Press.

Carey, S., (1985), *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Chen P., (1997), 'From Covariation to Causation: A Causal Power Theory', *Psychological Review*.

Dunbar, K. (1994), How Scientists Really Reason: Scientific Reasoning in Real-world Laboratories, in R. J. Sternberg & J. Davidson (eds.), *Mechanisms of insight*. Cambridge, MA: MIT Press.

Esbensen, B. M., Taylor, M., and Stoess, C. (1997), 'Children's Behavioral Understanding of Knowledge Acquisition'. *Cognitive Development 12* 53–84.

Glymour, C., (1998), 'Learning Causes', *Minds and Machines*, this issue, pp. 39–60.

Gopnik, A., (1996), "The Scientist as Child", *Philosophy of Science* 63, pp. 485–514.

Gopnik, A., (1998) 'Explanation as Orgasm' *Minds and Machines*, this issue, pp. 101–118.

Gopnik, A. and Wellman, H. M. (1994). The Theory Theory. in L.A. Hirschfeld and S.A. Gelman (eds.), *Mapping the Mind: Domain Specificity in cognition and culture*, pp. 257–293. Cambridge: Cambridge University Press.

Grice, H.P., (1989), *Studies in the Ways of Words*. Cambridge, MA: Harvard University Press.

Hempel, C.G., (1965), *Aspects of Scientific Explanation*. New York: Free Press.

Hutchins, E., (1995), *Cognition in the Wild*, Carnbridge, MA: MIT Press.

Kahneman, D., and Tversky, A. (1973), 'On the psychology of prediction', *Psychological Review* 80, pp. 237–251.

Keil, F. C., and Smith, W. C. (1996), Is there a Different "Basic" Level for Causal Relations? *Paper Presented at the 37th Annual Meeting of the Psychonomic Society* (November), Chicago, IL.

Keil, F., Smith, C., Simons, D., and Levin, D. 'Two Dogmas of Conceptual Empiricism'. *Cognition* (in press).

Keil, F. C. (1995), 'The Growth of Causal Understandings of Natural Kinds', in D. Sperber, D. Premack, and A. Premack (eds.), *Causal Cognition: A Multidisciplinary Debate*. Oxford: Oxford University Press.

Kitcher, P., (1989), 'Explanatory Unification and the Structure of the World', in P. Kitcher and W. Salmon (eds.), *Scientific Explanation*. Minneapolis: University of Minnesota Press.

Kitcher, P., (1992), 'The Naturalists Return', *Philosophical Review* 101, pp. 53–114.

Kitcher, P., (1993), *The Advancement of Science*. Oxford: Oxford University Press.

Leslie, A. (1995), 'A Theory of Agency', in A. L. Premack, D. Premack and D. Sperber (eds.), *Causal Cognition: A Multi-disciplinary Debate*, New York: Oxford, pp. 121–141.

Lindberg, D. C. (1992). *The Beginnings of Western Science: The European Scientific Tradition in Philosophical Religious and Institutional Context 600 B.C. – A.D. 1450*. Chicago: University of Chicago.

Locke, John, (1690), *An Essay Concerning Human Understanding*. New York: Dover. pp. 1959.

Magner, L.N. (1992), A History of Medicine. New York: Marcel Dekker, Inc..

Maffi, L. (1994), 'A Linguistic Analysis of Tzeltal Maya Ethnosymptomatology'. *Dissertation Abstracts International*, 55, pp. 950–901.
Millikan, R.G. A Common Structure for Concepts of Individuals, Stuffs, and Real Kinds: More Mama, More Milk and More Mouse. *Behavioral and Brain Sciences*. (in press).
Murphy, G.L. and Medin, D. (1985). 'The role of theories in conceptual coherence'. *Psychological Review* 92, pp. 289–316.
Nagel, E., (1961), *The Structure of Science*. (2nd edition), Indianapolis: Hackett, 1979.
Proffitt, D.R., and Gilden, D.L. (1989), 'Understanding Natural Dynamics', *Journal of Experimental Psychology Human Perception & Performance*, 15(2), pp. 384–393.
Putnam, H. (1975). 'The meaning of meaning', in his *Mind, Language and Reality*. London: Cambridge University Press.
Reichenbach, H., (1938), *Experience and Prediction*. Chicago: University of Chicago Press.
Salmon, W., (1989), 'Four Decades of Scientific Explanation', in P. Kitcher and W. Salmon (eds), Scientific Explanation. Minneapolis: University of Minnesota Press.
Scarborough, J. (1969), *Roman Medicine*. Ithaca, New York: Cornell University Press.
Simon, H., (1969), *The Sciences of the Artificial*. Cambridge, MA: MIT Press.
Sloman, S., Love, B., and Ahn, W., 'Mutability of features'. *Cognitive Science*. in press.
Spelke, E. (1994), 'Initial Knowledge: Six Suggestions', *Cognition*, 50, pp. 431–445.
Taylor, M., Esbenson, B. M., and Bennett, R.T. (1994), 'Children's Understanding of Knowledge Acquisition: The Tendency for Children to Report they have Always Known what they have just Learned'. *Child Development*, 65, pp. 1581–1604.
Thagard, P., (1997), 'Explaining Disease: Correlations, Causes, and Mechanisms', *Mind and Machines*, (this volume).
van Fraasen, B., (1977), 'The Pragmatics of Explanation', reprinted in R. Boyd, P. Gasper, and J.D. Trout (eds.), *The Philosophy of Science*. Cambridge, MA: MIT Press, 1991.
Waldrop, W. M. (1992), *Complexity: The Emerging Science a the Edge of Order and Chaos*. New York: Simon and Schuster.
Wason, P.C. (1968). 'Reasoning About a Rule', *The Quarterly Journal of Experimental Psychology*, 20, pp. 273–281.
Wason, P. and Johnson-Laird, P. (1972). *Psychology of Reasoning: Structure and Content*. London: Batsford.
Wellman, H., (1990), *The Child's Theory of Mind*. Cambridge, MA: MIT Press.
Wellman, H. M., & Gelman, S. A. 'Knowledge Acquisition in Foundational Domains', in D. Kuhn and R. Siegler (eds.), *Cognition, perception and language Vol 2. of Handbook of Child Psychology 5th ed.)*, New York: Wiley. (in press).
Wilson, R.A., (1994), 'Causal Depth, Theoretical Appropriateness, and Individualism in Psychology', *Philosophy of Science* 61, pp. 55–75.
Wilson, R.A., (1995), *Cartesian Psychology and Physical Minds*, New York: Cambridge University Press.
Wilson, R.A., (1998), 'The Mind Beyond Itself', in D. Sperber (ed.), *Metarepresentation*. Oxford: Oxford University Press.
Wilson, R.A., 'On the Realization of Mental Properties: Constitution, Determination, and Width', typescript.
Zola-Morgan, S. and Squire, L.R. (1993). Neuroanatomy of memory. Annual Review of Neuroscience 16, pp. 547–563.