The Ontic Account of Scientific Explanation.

Carl F. Craver[1]

According to one large family of views, scientific explanations essentially subsume a phenomenon (or its description) under a general representation (see Hempel 1965; Kitcher 1989; Churchland 1989; Bechtel and Abrahamsen 2005; Machamer, Darden, and Craver 2000). Authors disagree about the precise form that these representations should take: For Carl Hempel, they are generalizations in first order logic; for Philip Kitcher they are argument schemas; for Bechtel and Abrahamsen they are mental models; for Churchland, they are prototype vectors; for Machamer, Darden and Craver, they are mechanism schemas. Here, my focus is on the basic assumption that the philosophical dispute about scientific explanation is, or should be, about the representational form that such explanations take. While this *representational subsumption view* (RSV), in all of its guises, will likely be part of any theory of human understanding, the RSV is precisely the wrong place to begin developing a philosophical theory of explanation. Or so I shall argue.

Two philosophical objectives have been central the philosophical debate over the nature of scientific explanation for over 50 years. The first is *explanatory demarcation*: the theory should distinguish explanation from other forms of scientific achievement. Explanation is one among many kinds of scientific success; others include control, description, measurement, prediction, and taxonomy. A theory of explanation should say

[1] I thank Andreas Hütteman, Marie Kaiser, Alex Reutlinger and other members of the philosophy community at the Universtität zu Köln for support and discussion during the writing of this paper. I also thank Ken Aizawa, Kevin Amidan, Justin Garson, and Jim Tabery for feedback on earlier drafts. This paper was delivered at Duke University, and I am grateful to Robert Brandon, Andrew Janiak, Karen Neander, Alex Rosenberg, and Walter Sinnott-Armstrong for helpful comments.

how explanatory knowledge differs from these others and should say in virtue of what particular kinds of knowledge count as explanatory. The second goal is *explanatory normativity*. The theory should illuminate the criteria that distinguish good explanations from bad. The term "explanation" should not be an empty honorific; the title should be earned. A philosophical theory of explanation should say when the title is earned. My claim is that in order to satisfy these two objectives, one must look beyond representational structures to the ontic structures in the world. Representational subsumption, in other words, is insufficient as an account of scientific explanation. The fundamental philosophical dispute is ontic: it concerns the kinds of ontic structure that ought to populate our explanatory texts, whatever their representational format.

Some caveats will hopefully prevent misunderstandings. First, I do not claim that one can satisfy all of the normative criteria on explanatory models, texts, or communicative acts by focusing on ontic explanations alone. Clearly there are questions about how one ought to draw diagrams, organize lectures, and build elegant and useable models that cannot be answered by appeal to the ontic structures themselves. The ontic explanatory structures are in many cases too complex, reticulate, and laden with obfuscating detail to be communicated directly. Scientific explanations are constructed and communicated by limited cognitive agents with particular pragmatic orientations. These topics are interesting, but they are downstream from discussions of what counts as an explanation for something else. Our abstract and idealized representations count as conveying explanatory information in virtue of the fact that they represent certain kinds of ontic structures (and not others). Second, my topic is independent of psychological questions about the kinds of explanation that human cognitive agents tend to produce or

tend to accept. Clearly people often accept as explanations a great many things that they should reject as such. And people in different cultures might have different criteria for accepting or rejecting explanations. These facts (if they are facts) would be fascinating to anthropologists, psychologists, and sociologists. But they are not relevant to the philosophical problem of stating when a scientific explanation ought to be accepted as such. In the view defended here, scientific explanation is a distinctive kind of achievement that cultures and individuals have to learn to make. Individual explanatory judgments, or cultural trends in such, are not data to be honored by a normative theory that seeks to specify when such judgments go right and when they go wrong. Finally, I do not suppose that there is one and only one form of scientific explanation. Though at times I adopt a specifically causal-mechanical view of explanation (see Craver 2007), and so will describe the ontic structures involved in explanation as causal or mechanistic, I intend the term *ontic structure* to be understood much more broadly. Other forms of ontic structure might include: attractors, final causes, laws, norms, reasons, statistical relevance relations, symmetries, and transmissions of marks, to name a few. The philosophical dispute about explanation, from this ontic perspective, is about which kinds of ontic structure properly count as explanatory and which do not.

I proceed as follows. In Section 1, I disambiguate four ways of talking about explanation: as a communicative act, as a representation or text, as a cognitive act, and as an objective structure in the world. The goals of that discussion are to distinguish these senses of explanation and to highlight some distinctive conceptual contributions that the ontic conception makes to our speaking and thinking about explanations. In Section 2, I illustrate how appeal to ontic explanations is essential for marking several crucial

normative dimensions by which scientific explanations are and ought to be evaluated: the distinction between how-possibly and how-actually-enough explanations, the distinction between phenomenal descriptions and explanations, the difference between predictive and explanatory models, and the requirement that explanatory models should include all and only information that is explanatorily relevant to the phenomenon one seeks to explain. In Section 3, I review how these normative dimensions long ago raised problems for Hempel's covering-law model, the once-dominant idea that explanations are arguments (texts) with a description of the *explanandum* phenomenon as their conclusion. In Section 4, I use Churchland's PDP model of explanation as an exemplar of psychologistic theories to illustrate how cognitivist models of explanation presuppose, rather than satisfy, the normative distinctions laid out in Section 3. In Section 5, show how the ontic conception provides a satisfyingly simple answer to the question: How can a false representation explain?

1. The Ambiguities of "Explanation"

Consider four common modes in which people (including scientists) talk about explanation. Suppose, thinking of one's neuroscience professor, one says:

(S1) Jon explains the action potential (Communicative Mode).

One might imagine Jon in front of a classroom, writing the Hodgkin and Huxley model of the action potential on a chalkboard. Alternatively, we might imagine him writing a textbook that walks, step-by-step, through the complex mechanisms that give rise to action potentials. Explanation, so understood, is a communicative act. It involves an explainer, an audience, and a text (a lecture or book, in this case) that conveys information from the explainer to the audience. If everything goes right, Jon manages in

his lecture to convey information about action potentials to an audience, and the audience comes to understand how action potentials are produced.

Explanatory communications of this sort might fail in at least three ways. First, Jon might successfully deliver a false explanation. He might explain (incorrectly) that action potentials are produced by black holes in the endoplasmic reticulum. We can imagine excited students understanding Jon's lecture and dutifully reporting it back on the exam. Jon explains the action potential to the class (that is, he gave them a model of action-potential generation), but the explanation is false. In a second kind of failure case, Jon unsuccessfully delivers a true explanation. He might, for example, give an impeccably accurate lecture about the action potential but leave his students completely confused. The lecture fails, we might suppose, because it presupposes background knowledge the students lack, or because it is delivered in a language the students are unprepared to handle. Finally, we might imagine Jon delivering an impeccably organized and conversationally appropriate lecture to undergraduate students who, because they are distracted by other plans, fail to understand what Jon is telling them. The explanation fails as a communicative act, but it is not Jon's fault. His audience just didn't get it.

In contrast to this communicative mode, we sometimes talk about explanation in the *ontic mode*, as a relation among features of the world. One says, for example that:

> (S2) The flux of sodium (Na+) and potassium (K+) ions across the neuronal membrane explains the action potential. (Ontic Mode)

S2 is not at all like S1. In cases like S2, the items in the subject position are not intentional creatures (like Jon); they are states of affairs. And no text about a topic is transmitted from an explainer to an audience. The explanatory relation described in S2 is

not properly fleshed out in terms of the delivery of information via a text to an audience. There is no text, no representation, no information (in the colloquial sense).[2] It would appear, in fact, that S2 could be true even if no intentional creature knows or ever knew the fact that S2 expresses. The term "explains" in S2 is synonymous with a description of the kinds of factors and relations (ontic structures) that are properly taken to be explanatory; as noted above, examples include causes, causal relevance, components, laws, and statistical relevance relations. Wesley Salmon expressed precisely this contrast as follows:

> The linguistic entities that are called 'explanations' are statements reporting the actual explanation. Explanations, in this [ontic] view, are fully objective and, where explanations of nonhuman facts are concerned, they exist whether or not anyone ever discovers or describes them. Explanations are not epistemically relativized, nor (outside of the realm of human psychology) do they have psychological components, nor do they have pragmatic dimensions. (1989, 133)

Salmon credits Coffa (1974) with this insight and notes that even Hempel, at times, could be read as embracing the view that laws themselves (rather than law statements or generalizations, which are representations of laws) provide ontic explanations for explanadum events and regularities. As Coffa explains, Hempel's deductive-nomological formulation of the covering law model is susceptible of either an ontic or epistemic interpretation. However, his inductive-statistical formulation has, at bottom, an

[2] This claim is bracketed to non-intentional contexts, those in which the explanation in question does not appeal to the flow of information, as do some explanation in genetics, neuroscience, psychology, and the social sciences. In such cases, of course, the information is not passed from the speaker of the explanation to its audience, but rather among the components in the explanation.

irreducible epistemic component. This is because Hempel defines the relevant probabilities in such explanations relative to the presumed background knowledge of the scientists. For Coffa, the need to relativize what counts as an explanation to what people know or believe was a major strike against the account. In his view, "no characterization of inductive explanation incorporating that feature [epistemic relativization] can be backed by a coherent and intelligible philosophy of explanation" (1974, 57). The ontic mode captures this objective way of talking about explanation.[3]

It is worth emphasizing that the term "explanation" in S1 and S2 is ambiguous, as revealed by the inability to meaningfully combine the two sentences into one, as in:

(S1+2) Jon and the flux of Na+ and K+ ions across the neuronal membrane explain the action potential.

If we think of explanation primarily in the communicative mode, (S1+2) appears odd because Na+ and K+ ions don't deliver lectures or produce diagrams with the intention of delivering information to an audience. If we think of explanation in the ontic mode, as, for example, a matter of producing, constituting, or otherwise being responsible for the *explanandum* phenomenon, then (S1+2) appears odd because it appears to assert that Jon causes, produces, or otherwise is responsible for the generation of action potentials (generally), which is clearly false. For these reasons, it would be a kind of conceptual mistake to think that an analysis of "explanation" in the sense expressed in S1 could serve as an analysis of "explanation" in the sense expressed in S2. This is not to say the two are unrelated. In particular, whether Jon has provided his class with a correct

[3] The German verb "erklären" is not ambiguous like the English word "explanation." The verb contains the idea of "making clear," which automatically suggests the communicative or representational mode.

explanation of the action potential would seem to depend on whether Jon's lecture correctly indicates how action potentials are produced or constituted. The endoplasmic black hole hypothesis makes this clear. That is, whether or not Jon's explanatory communicative act (described in S1) fails in the first sense described above will depend on whether his text matches (to a tolerable degree) the patterns of causation, constitution, and responsibility that *in fact* explain the production of action potential (as described in S2).

To explore this connection a bit further, consider a third mode of thinking about explanations. Where S1 places Jon, the communicative agent, in the subject position, and S2 places worldly states of affairs in the subject position, this third way of speaking puts Jon's explanatory text in the subject position:

> (S3) The Hodgkin-Huxley (HH) model explains the action potential. (Textual Mode.)

The HH model of the action potential, one of the premier theoretical achievements in the history of neuroscience, is a mathematical model that describes how the membrane voltage of a neuron changes as a function of ionic conductances and how ionic conductances change as a function of voltage and time. My point does not turn on the fact that a specifically *mathematical* model appears in the subject position; rather, S3 is meant to apply generally to any text: it might be an article, book, cartoon, diagram, film, graph, or a lecture. A text, in this sense, is a vehicle for conveying intentional content from a communicator to an audience. Hodgkin and Huxley communicated their understanding of the current-voltage relations in neuronal membranes to the rest of us in the form of a

compact mathematical representation from which we (the audience) might extract a wealth of pertinent information about this topic.

Yet it would be a mistake to put John and the HH model together as the conjoined subjects of a sentence such as

(S1+3) Jon and the HH model explain the action potential

because they explain the action potential in different ways: Jon as a communicative agent, and the HH model as a communicative text. It would not be a confusion of this sort to assert that Jon, Hodgkin, and Huxley explained the action potential to the class (perhaps Jon invited some illustrious guests). Nor would it be confused to claim that Jon's lecture, the equivalent circuit diagram, and the HH model explained the action potential. In these last two sentences, the term "explanation" applies univocally to the three objects listed in the subject position.

It would be a confusion, however, to put the HH model and ionic fluxes together as the conjoined subjects of a sentence such as:

(S2+3) Ionic fluxes and the HH model explain the action potential

for they again explain the action potential in different ways: ionic fluxes, as states of affairs that produce or constitute action potentials, and the HH model as a communicative text. Equations do not produce action potentials, though action potentials and their mechanisms can be described using equations. The HH model might be included in the explanatory text, but the equation is neither a cause nor a constituent of action potentials. This confusion is propagated by those who think of the HH model as a "law" that "governs" the action potential (e.g. Weber 2005) rather than as a mathematical generalization that describes how some of the components in the action potential

mechanism behave (see Bogen 2005; Craver 2005; Craver 2007). To put the point the other way around, it would be wrong to claim that Jon used ionic fluxes to explain action potentials to the class (unless, for example, he were to illustrate the process of diffusion by placing dye in the bottom of a beaker, in which case the demonstration becomes a "text" that is intended to convey information to a class).

Finally, let us consider a more mentalistic way of speaking about explanation (or, less awkwardly, about understanding). We might think that a cognitive agent explains/understands a phenomenon by activating a mental model that in some sense fits the phenomenon to be explained. Churchland speaks of explanations, as I discuss below, as involving the activation of a prototype in the connectionist networks of one's brain. Similarly, Bechtel and Abrahamson insist that explanation is "essentially a cognitive activity". Directly contrary to the reading in S2, they claim that what figures in explanation is not "the mechanisms in the world" but "representations of them" (Bechtel and Abrahamson 2005, 425). To express this very reasonable thought, we should recognize a fourth way of speaking:

> (S4) Jon's mental representation of the mechanism of the action potential explains the action potential (Cognitive Mode).

S4 is no doubt a bit strained to the native English-speaker's ear. Typically, in these situations, we would speak not of explanation but of understanding. Jon, in this case, *understands* the action potential when Jon can activate a mental representation of the requisite sort and, for example, answer questions about how the action potential might differ depending on different changes in background conditions, ion concentrations, distributions of ion channels, cell morphology, and the like. But let us put this worry

aside until the next section in order to draw out some important differences between S4 and the others.

The subject position of S4 is occupied by a mental representation. The subject is not a cognitive agent but, as it were, a part or sub-process of the agent's cognitive architecture. The mental representation itself has no communicative intentions of the sort that Jon has. And it is hard to make sense of the idea that such a representation has an audience with which it is attempting to communicate. Though mental representations are said to influence one another, subsume one another, and the like, it would be an illegitimately homuncular sort of thought to say that, for example, one mental representation understands the other. There is nobody "in John's head" to read the representation and understand it. And it would be wrong to say that the mental representation explained something to Jon (in the sense of S1), since Jon, as the possessor of the mental representation, already understands it quite well. Thus it seems to me something of a category mistake to assert that:

(S1+4) Jon and his mental representations explained the action potential.

Though it is at least plausible to say that Jon is able to explain the action potential to the class in virtue of his having a set of stored mental representations about action potentials, the mental representations and Jon explain in different ways. Likewise, for reasons we've already discussed, it would be a mistake to assert that:

(S2+4) Jon's mental representations and ion fluxes explain the action potential,

Jon's mental representations do not cause or produce action potentials (unless they drive him to do some electrophysiological experiments). And although ionic fluxes are certainly involved in the production of Jon's mental representations (given that such

representations must be implemented somehow in neural architectures), the ionic fluxes in Jon's brain do not subsume action potentials, as do his abstract representations of the action potential mechanisms. S4 is clearly closest to the textual reading of explanation statements:

> (S3+4) Jon's mental representations and the HH model explain the action potential.

Rightly or wrongly, many think of mental representations as texts or images written in the mind or in our neural architectures. If so, it is easy to think that mental representations and scientific representations might play the same kind of role and might be involved in the same kind of explanatory process. There are important differences, however, between these two kinds of explanation. First, the HH model might explain the action potential even if Jon never learns it. In S4 we are concerned with cognitive achievements of a single mind, not with the explanatory advance of a science (as appears to be the concern of S3). The HH model "covers" or "subsumes" many features of the action potential regardless of whether Jon ever hears about it. For the HH model to be relevant to Jon's understanding of the action potential, let us allow, he has to form an internal representation of something like the HH model and activate it. But the HH model itself does not need to be "activated" to count as an explanation. Even if (implausibly enough) there is a brief moment in time when nobody in the world is thinking about the HH model in relationship to action potentials, there remains a sense in which the HH model continues to explain the action potential during our cognitive slumbers (if, indeed, the HH model is an explanation of the action potential; a topic to which I return below).

The simple point is that the term "explanation" has four common uses in colloquial English: 1) to refer to a *communicative act*, 2) to refer to a cause or a factor that is otherwise responsible for a phenomenon (the *ontic* reading), 3) to refer to a *text* that communicates explanatory information, and 4) to refer to a *cognitive* act of bringing a representation to bear upon some mysterious phenomenon. These uses are no doubt related. Explainers (we might suppose) understand a phenomenon in virtue of having certain *cognitive* representations, and they use explanatory *texts* (such as the HH model) to represent *ontic* explanations (such as the production of action potentials by ionic fluxes) in order to *communicate* that understanding to an audience. Though these senses of "explanation" are subtly related to one another, they are not so subtly *different* senses of explanation. It would be a mistake to conflate them.

In particular, there is an especially clear line between S2, on the one hand, and S1, S3, and S4 on the other. S1, S3, and S4 each depend in some way on the existence of intentional agents who produce, interpret, manipulate, and communicate explanatory texts. Jon's communicative act of explanation presupposes a communicator and an audience. The HH model is a scientific text produced, learned, and applied by intentional agents in the act of discovering, explaining, and understanding action potentials. It is a called a model in part because it is a representation that intentional creatures can use for the purposes of making inferences about a worldly system. And Jon's internal representation, or mental image, is likewise dependent for its existence on Jon's being the kind of creature that thinks about things. I suppose it is possible to understand the term "model" in a more technical, logical, or set-theoretic sense, but even in this technical

reading the notion depends for its existence on creatures that are able to, for example, form inferences and apply general frameworks in specific instances.

S2, the ontic mode of thinking about explanation, does not depend on the existence of intentional agents in this way. A given ontic structure might cause, produce, or otherwise be responsible for a phenomenon even if no intentional agent ever discovers as much. This ontic way of talking about explanation allows us to express a number of reasonable sentences that would be strained, if not literal nonsense, if our thinking about explanation were tied to the modes expressed in S1, S3, and S4. Here are some examples:

(A) Our world contains undiscovered phenomena that have explanations.

(B) There are known phenomena that we cannot currently explain (in the sense of S1, S3, or S4) but that nonetheless have explanations.

(C) A goal of science is to discover the explanations for diverse phenomena.

(D) Some phenomena in our world are so complex that we will never understand them or model them, but they have explanations nonetheless.

If we tie our thinking about explanation to the existence of creatures that are able to represent, communicate, and understand phenomena, each of these sentences is awkward or nonsensical. If one allows for an ontic way of thinking about explanation, however, each of these sentences is relatively straightforward and non-elliptical. (A) concerns aspects of the world that nobody has ever represented or that nobody ever will represent. If explanation requires representation by an intentional agent, then this should not be possible. B, C, and D also recognize a distinction between whether or not a phenomenon has an explanation, on the one hand, and whether anyone knows or can otherwise construct the explanation for it, on the other. A-D are very natural things to say.

More importantly, A-D indicate an asymmetric direction of fit between the representation-involving ways of talking about explanation and the ontic mode. In particular, it would appear that the adequacy of our communicative acts, our scientific texts, and our mental models depends in part on whether they correctly inform us about the features of the world that cause, produce, or are otherwise responsible for the phenomena we seek to explain. While Jon might be able to convey his endoplasmic black-hole model of the action potential to his students (and thus to explain his model to them) Jon would not thereby explain the action potential to them. His putative explanation would merely leave them confused, whether they know it or not. If we treat this black hole model as one of Jon's mental representations activated when he thinks of action potentials, then it seems right to say that although Jon thinks he understands the action potential, he is deeply mistaken about this; in fact, he has only the illusion of understanding the action potential. And the same can be said of false models; they might vary considerably in the accuracy with which they describe the explanation for the action potential. If the philosophical topic of explanation is to provide criteria of adequacy for scientific explanations, then the ontic conception is indispensable: explanatory communications, texts, and representations are evaluated in part by the extent to which they deliver more or less accurate information about the ontic explanation for the *explanandum* phenomenon.

2. Adequate Explanations and the Ontic Conception. In many areas of science, explanatory texts are taken to be adequate to the extent that they correctly describe the causes (etiological explanations) or the underlying mechanisms (constitutive explanations) responsible for the phenomenon one seeks to explain (the *explanandum*

phenomenon) (See Machamer Darden and Craver2000; Craver 2007; Bechtel and Abrahamsen 2005). In such areas of science, successful models contain variables that stand for causally relevant properties or features of the system and represent the appropriate relations among those variables. Successful communication of explanatory information (as opposed to misinformation) conveys information about those causally relevant features and their relations. And finally, one understands (rather than misunderstands) the *explanandum* phenomenon to the extent that one correctly grasps the causal structure of the system at hand.

The importance of truth to scientific explanation generally is recognized in the commonplace distinction between a *how-possibly model* and a *how-actually* model (Dray 1957; MDC 2000). Gastric ulcers might have been caused by emotional stress (as it was once thought), but they are in fact caused by *Helicobacter pylori* bacteria (see Thagard 1999). Action potentials might have been produced by a distinctive form of animal electricity, but they are in fact produced by fluxes of ions across the cell membrane. The earth might have been at the center of the solar system with the moon, sun, and planets revolving around it, but it is not. One might form elegant models describing these putative causes and constitutive mechanisms, and one might use such models to predict various features of the *explanandum* phenomenon, and such models might provide one with the illusion that one understands how an effect is brought about or how a mechanism works. However, there is a further fact concerning whether a plausible explanation is in fact the explanation.

To claim that truth is an essential criterion for the adequacy of our explanations, one need not deny that paradigmatically successful explanatory models explicitly make

false assumptions or presume operating conditions that are never seen in reality. An explanatory model in physics might assume that a box is sliding on a frictionless plane. An explanatory model in electrophysiology might presume that an axon is a perfect cylinder or that the membrane obeys Ohm's law. A physiologist might model a system in a "wild-type" organism by presuming that all individual organisms in the wild type are identical. Such idealization is often required in order for one to form a parsimonious yet general description of a wide class of systems. Yet this undeniable fact about scientific models need not lead one to abandon the not-so-subtle difference between models that incorrectly describe how something might have worked from those that describe, more accurately, how it in fact works. In other words, whatever we want to say about idealization in science, it should not lead us to the conclusion that there is no explanatory difference between a model that describes action potentials as being produced by ionic fluxes, on the one hand, and one that describes it as being produced by black holes, on the other. Perhaps then, the appropriate distinction is not between how-possibly and how-actually, but between how-possibly and how-actually-within-the-limits-of-idealization.

Yet my point about the centrality of the ontic conception to our criteria of explanatory adequacy goes beyond the mere claim that our explanations should be true (or approximately true). Not all true models are explanatory. Models can be used to describe phenomena, to summarize data, to calculate undetected quantities, and to generate predictions (see Bogen 2005). Models can play any or all of these roles without explaining anything. Models can fall short as explanations because: (1) they are purely descriptive or *phenomenal models*; (2) they are purely *predictive models*; (3) they are mere *sketches* of the components and activities of a mechanism with gaps and question-

marks that make the explanation incomplete; or (4) because the model includes explanatorily irrelevant factors. Consider these in turn.

(1) *Phenomenal Models.* Scientists commonly draw a distinction between models that merely describe a phenomenon and models that explain it. Neuroscientists such as Dayan and Abbott, for example, distinguish between purely descriptive mathematical models, models that "summarize data compactly," and mechanistic models, models that "address the question of how nervous systems operate on the basis of known anatomy, physiology, and circuitry" (2001, xiii). Mechanistic models describe the relevant causes and mechanisms in the system under investigation. The distinction between purely descriptive, phenomenal models and mechanistic models is familiar in many sciences. Snell's law describes how light refracts as it passes from one medium to another, but the law does not explain why the path of light changes as it does. To explain this principle, one must appeal to facts about how light propagates or about the nature of electromagnetic phenomena. (Of course, one might explain the angle of refraction of a beam of light by appeal to the fact that the light crossed between two media in which it has different velocities. However, we are interested here in explaining why light generally bends when it passes from one medium to the next. Snell's law tells us that our beam of light is not alone in exhibiting this mysterious behavior, but it does not tell us why light generally behaves this way.)

Precisely the same issue arose with respect to the HH model of the action potential. As part of building their "total current equation," Hodgkin and Huxley (1952) generated equations to model how the conductance of a neuronal membrane to sodium and potassium changes as a function of voltage during an action potential. The equations are

surprisingly accurate (approximately true), but they leave it utterly mysterious just how the membrane changes its conductance during an action potential. Hodgkin and Huxley are explicit about this explanatory limitation in their model. To explain these conductance changes, scientists needed first to discover the membrane-spanning channels that open and close as a function of voltage (Bogen 2005, 2008; Craver 2006, 2007, 2008; Hille 2001). The signature of a phenomenal model is that it describes the behavior of the target system without describing the ontic structures that give rise to that phenomenon.

(2) *Purely predictive models.* Explanatory models often allow one to make true predictions about the behavior of a system. Indeed, some scientists seem to require that explanatory models must make new predictions. Yet not all predictively adequate models are explanatory. A model might relate one effect of a *common cause* to another of its effects. For example, one might build a model that predicts the electrical activity of one neuron, A, on the basis of the activity of another neuron, B, when the activities of both A and B are in fact explained by the activity in a "parent" neuron, C, that synapses onto both A and B (while A and B have no influence on each other). A model might relate *effect to cause.* One might, that is, build a model that predicts the behavior of neuron C in the above example on the basis of the behavior of neurons A or B. One can infer that a brain region is active on the basis of changes in the ratio of oxygenated to de-oxygenated hemoglobin in the vasculature of that brain region. This law-like correlation makes functional magnetic resonance imaging of the brain possible. Yet nobody to my knowledge believes that the changes in oxygenation explain neuronal activity in these brain regions. The explanation runs the other way around; changes in neural activation cause (and so explain) changes in regional blood flow. Finally, a model might relate two

events that follow one another in a regular *sequence* but that, in fact, have no explanatory connection. One can predict that the ballgame will begin from the performance of the national anthem, but the performance of the national anthem does not explain the start of the game. The point of these examples is that models may lead one to expect a phenomenon without thereby explaining the phenomenon. These judgments of scientific common sense seem to turn on the hidden premise that explanations correctly identify features of the ontic structures that produce, underlie, or otherwise responsible for the *explanandum* phenomenon (see Salmon 1984). Expectation alone does not suffice for explanation.

(3) *Sketches.* A third dimension for the evaluation of scientific models is the amount of detail that they provide about the causal structure of the system in question. A model might "cover" the behavior of a system at many grains of description. It might be a phenomenal model, as described above, in which case it serves merely as a description, rather than an explanation, of the system's behavior. At the other end of the spectrum, it might supply a fully worked out description of all of the components, their precise properties, their precise spatial and temporal organization, all of the background and boundary conditions, and so on. It is rare indeed that science achieves that level of detail about a given system, in part because a central goal of science is to achieve generalization, and such particularized descriptions foil our efforts to build generalizable models. Between these poles lies a continuum of grains of detail. A mechanism sketch is a model of a mechanism that contains crucial black boxes or filler-terms that, at the moment, cannot be filled in with further details. For example, one might sketch a model of memory systems as involving encoding, storage, and retrieval without having any

precise ideas about just how memories are encoded in the brain, how or where they are stored, or what precisely it would mean to retrieve them. Such a sketch might be true, or approximately true, and nonetheless explanatorily shallow. One can deepen the explanation by opening these black boxes and revealing their internal causal structure. In doing so, one allows oneself to answer a broader range of questions about how the phenomenon would differ were one or the other feature of the mechanism changed (cf. Woodward 2003). This ability is typically taken to be an indirect measure of one's understanding of how the system works. The crucial point about sketches for present purposes is that the spectrum from phenomenal model, to sketch, to schema, to fully instantiated mechanism is defined by the extent to which the model reveals the precise details about the ontic explanation for the phenomenon. Again, it would appear, the ontic explanation plays an asymmetric and fundamental role in our criteria for assessing explanations.

(4) *Relevance.* An explanatory text for a given phenomenon ought to include all and only the factors that are explanatorily relevant to the *explanandum* phenomenon. While it is true that people with yellow fingers often get lung cancer, the yellow fingers are explanatorily irrelevant to the lung cancer. A putatively explanatory model that included finger color as part of the explanation for Nancy's lung cancer would be a deeply flawed explanatory model.

As discussed in the previous section, explanations are sometimes spoken of as communicative acts, texts (e.g., models), and representations. So conceived, explanations are the kinds of things that can be more or less complete and more or less accurate. They might include more or less of the explanatorily relevant information. They might be more

or less deep. Conceived ontically, however, the term explanation refers to an objective portion of the causal structure of the world, to the set of factors that produce, underlie or are otherwise responsible for a phenomenon. Ontic explanations are not texts; they are full-bodied things. They are not true or false. They are not more or less abstract. They are not more or less complete. They consist in all and only the relevant features of the mechanism in question. There is no question of ontic explanations being "right" or "wrong," or "good" or "bad." They just are.

The point of this section is that ontic explanations, the ontic structures in the world, make an essential contribution to the criteria for evaluating explanatory communications, texts (models), and mental models. Good mechanistic explanatory models are good in part because they correctly represent objective explanations. Mere how-possibly models describe the wrong causes or wrong mechanisms whereas how actually models get it right. Phenomenal models describe the phenomenon without revealing the ontic structures that produce it. Merely predictive models describe correlations but not causal structures. Mechanism sketches leave out relevant portions of the causal structure of the world. The issue here is not merely that an explanation must be true: predictive models, phenomenal models, sketches, and models containing irrelevancies might be true but explanatorily inadequate. The ontic structure of the world thus makes an ineliminable contribution to our thinking about the goodness and badness of explanatory texts. The traditional philosophical problem of explanation was to provide a model that embodies the criteria of adequacy for sorting good explanations from bad. One cannot solve that problem without taking the ontic aspect of explanation seriously.

Let me put this another way: the norms of scientific explanation fall out of a prior commitment on the part of scientific investigators to describe the relevant ontic structures in the world. Explanation, in other words, is intimately related to the other aspects of science, such as discovery and testing. The methods that scientists use to discover how the world works, the standards to which they hold such tests, are intimately connected with the goal of science to reveal the ontic structures that explain why the phenomena of the world occur and why they occur as they do. One cannot carve off the practice of building explanations from these other endeavors. These methods and products of the scientific enterprise hang together once one recognizes that science is committed, *ab initio*, to giving a more or less precise characterization of the ontic structure of the world.

The commitment to realism embodied in these claims can be justified on several grounds. It is justified in part because it makes sense of scientific-commonsense judgments about the norms of explanation. It is also justified by reference to the fact that an explanation that contains more relevant detail about the responsible ontic structures are more likely, all things equal, to be able to answer more questions about how the system will behave in a variety of circumstances than is a model that does not aim at getting the ontic structures that underlie the phenomenon right. This follows from the fact that such models allow one to predict how the system will behave, for example, if parts of a mechanism are broken, changed, or rearranged and so how the mechanism is likely to behave if it is put in conditions that make a difference to the parts, their properties, or their organization. It is always possible (though never easy) to contrive a phenomenally adequate model post-hoc if and when the complete input-output behavior of a system is known. However, the critical question is how readily we can discover this input-output

mapping across the full range of input conditions without knowing anything about the underlying mechanism. We are far more likely to build predictively adequate models when aspects of the mechanism are known. Finally, models that reveal objective causal structures automatically reveal knobs and levers in the world that might be used for the purposes of bringing parts of it under our control (Woodward 2003).

To illustrate the importance of the ontic aspect of explanation for developing a philosophical theory of scientific explanation, I now consider two models of explanation that, at least on some readings, neglect the importance of the ontic mode. I argue that they fail to embody the criteria of adequacy for scientific explanations because they focus their attention on representations rather than on the ontic structures those representations represent.

3. *The CL model*. One systematic (though somewhat uncharitable) way of diagnosing the widely acknowledge failure of the CL model is to see it as emphasizing the explanatory representations over the ontic structures they represent. This is not the only, nor even the most familiar, diagnosis. Others (such as Churchland and Bechtel) argue that the CL model fails because it insists on formulating explanations in propositional logic. Such critics respond the shortcomings of the CL model by developing new representational frameworks that are more flexible and more cognitively realistic. If my diagnosis is correct, such revisions fail to address the core problems with the CL model as a theory of scientific explanation.

According to the CL model, explanations are arguments. The conclusion of the argument is a description of the *explanandum* phenomenon. The premises are law-statements, canonically represented as universal or statistical generalizations, and

descriptions of antecedent or boundary conditions. Explanation, on this view, is expectation: the explanatory argument shows that the description of the *explanandum* phenomenon follows, via an acceptable form of inference, from descriptions of the laws and conditions. In this sense, explanations show that the *explanandum* phenomenon was to be expected given the laws and the conditions. The emphasis is on the representational structures: *statements* of the laws, *descriptions* of the conditions, *entailment* relations, and human *expectations*.

I say that this characterization is somewhat uncharitable for two reasons. First, the CL model typically requires that the premises of the explanatory argument be true, i.e., that the law statements describe real laws and that the descriptions of conditions are accurate. Second, and more fundamentally, the logical force with which the *explanandum* statement follows from the premises might be taken to mirror the sense in which the *explanandum* phenomenon had to happen or was more likely to happen given the laws and the initial conditions. One might more charitably interpret Hempel as suggesting that that the inferential necessity in the argument mirrors or expresses the corresponding natural necessity in the world. And as Salmon (1989) pointed out, there are passages in Hempel's classic statement of the CL model that lend themselves to such an ontic interpretation: the laws, not law-statements, explain. Be that as it may, Hempel does not appear to have recognized this ambiguity in his own writing, and it is certainly in the keeping with the program of logical empiricism to think that all the essential features of science could be captured with the expressive formalism of logic. The commitment to

"natural necessity" in this putatively more charitable reading, in fact, does violence to Hempel's strongest empiricist convictions.[4]

Let me amplify a bit. On the most austerely empiricist interpretation of the CL model, it would be incorrect to say that the logical or inferential necessity of the argument "mirrors" a kind of natural necessity with which events follow the laws. According to this interpretation, the universal generalizations used to express universal laws are true summaries of events; they assert that all X's that are F are, as a matter of fact, also G. There is no further thing, the necessity of a law, that makes it the case that all Xs that are F are also G. Likewise, one might understand probabilistic laws as asserting objective frequencies. If we count up all of the X's that are F, we find as a matter of fact that some percentage of them are G. There need be no further fact that explains why G holds with this frequency in the population. In response to various counter-examples to the CL model, its defenders began to place more restrictions on the representations of laws. When it was objected that one could, according to this model, explain why a particular coin in Goodman's pocket is a dime on the basis of the claim that all the coins in Goodman's pocket are dimes, the response was to demand that laws make no reference to particulars, such as Goodman, or particular places, such as his pocket, or particular times, such as t= March 17, 1954 (see Ayer 1970). The formal structure of the representation, in other words, was called upon to block the counter-examples.

[4] As Ken Aizawa (personal communication) notes, the CL model arguably can accommodate sentences A-D of section 1. If one takes the CL model to equate explanation with rational expectability rather than rational expectation, then one can say that there are explanations to be discovered and explanations so complex that we will never know them.

But it appears that no amount of formal modification could block some very serious problems. In particular, the account could not satisfy the criteria of adequacy sketched in the previous section. First, the model does not, by itself, have machinery to distinguish phenomenal descriptions from explanations. Asked why a given X that is F is G (e.g., why a particular raven is black) the CL model famously appeals to the generalization that all Fs are Gs (e.g., all ravens are black). But one might reasonably object that such an explanation fails to discharge the request for explanation and, instead, merely lists the *explanandum* phenomenon as one of many phenomena, each of which is equally mysterious. Likewise, to explain why an action potential has a particular form, it does little to provide a generalized description of that form. One wants to know why action potentials have that form, not that all action potentials, in fact, have it. Clearly what one wants is an account of the ontic structures, in this case mechanisms, that give rise to action potentials (see Bogen 2005; Craver 2005). Perhaps one could describe such mechanisms in terms of a series of law statements about the internal causal structure of the action potential. Indeed, one might see the HH model as offering a sketch of just such an account. (I am not in favor of this way of talking, but will entertain it here to make my limited point.) My limited point is that there is a difference between such a mechanistic model, which reveals internal causal structures, and a phenomenal model, which simply generalizes the phenomenon; and crucially, the difference between them is not a formal difference but a difference in what is being described. The mechanistic description describes parts and processes at a lower level than the action potential, and this shift in levels is not a formal difference in the representation but an ontic difference between a whole (the action potential) and its parts. To mark the difference between a phenomenal

model and a mechanistic model (1 above), that is, one must appeal to the (quasi-mereological) structures of the world that relate the *explanans* to the *explanandum*.

Second, if one sticks with the austere empiricist reading of the CL model (that is, one that is not supplemented with some sort of ontic difference between laws and accidental generalizations), then the CL model does not recognize a distinction between generalizations that are explanatory and generalizations that are not. It should not matter whether two causally independent effects of a common cause explain one another or whether an effect explains its cause, or whether one type of event is explained by another type of event that always (or regularly) precedes it in time. That is, the CL model in its austere empiricist form does not distinguish explanatory models from merely predictive models (2 above). Indeed, the very idea that explanation is expectation would appear to insist that any predictive model is *ipso facto* an explanatory model. This is why the "prediction-explanation symmetry thesis" was heavily debated by proponents and opponents of the CL model alike. (Hempel abandoned it quickly.) The difference between explanation and prediction, which is fundamental to providing an adequate account of explanation, seems to rely not on some feature of the way we represent the world but rather on some feature of the world that distinguishes explanatory relations from mere correlations.

Third, the CL model in its austere form does not appear to mark a distinction between sketches and more complete descriptions of a mechanism. So long as the model suffices to derive a description of the *explanandum* phenomenon, it counts as an explanation (full stop). Grant that a defender could reconstruct multilevel explanation as one finds in neuroscience and physiology by describing, as it were, laws within laws all

the way down. That is not the issue. The issue before us is whether the CL model recognizes that by exploding black boxes and revealing internal causal structures one is, ipso facto, providing a deeper explanation. The model would become more and more complex, of course, as it includes more and more of the internal causal structure, but nothing in the formal structure of the model would indicate that the model was getting deeper. For that, one must appeal to features of the world that the model describes.

Finally, the CL model in its austere form does not recognize a difference between relevant and irrelevant explanatory factors. It is generally true that men who take birth control pills fail to get pregnant, and nothing in the formal structure of the CL model instructs us to jettison the irrelevant conjunct in the antecedent of this conditional. A similar problem arises for explanations of general laws. The predictive value of a model is unaffected (at least in many cases) by the inclusion of irrelevant detail. If one set of law statements and boundary conditions, K, entails another set, P, then the conjunction (K and S), where S is any arbitrary sentence that fails to contradict a member of K, also entails P. So the HH model plus Kepler's laws explains the HH model. Hempel called this the problem of irrelevant conjunction (Hempel 1965, 273, fn 33). This is a problem because it conflicts with the common scientific practice of filtering out irrelevant factors from explanations. A mechanistic explanatory model suffers, for example, if it includes irrelevant parts that are not in the mechanism, irrelevant properties that play no causal role, or irrelevant activities that are sterile in the mechanism. The important point for present purposes, however, is that it would appear that explanatory relevance is not a feature of the formal structure of an argument but rather of the kinds of ontic structures that the representation describes.

These kinds of objection to the CL model are by now thoroughly familiar to philosophers of science. What is less familiar, I suppose, is the thought that these problems require for their solution that one shift one's focus away from the representational structures of explanatory texts to features of the systems they represent. This is the insight of the ontic conception of explanation. The solution to these puzzles, and so the fundamental tasks of providing a philosophical account of explanation, is not to be discovered by building elaborate theories about how explanatory information is represented. Though the question of how such information is or ought to be represented is interesting and worthwhile, it will not by itself answer the questions that a narrower, normative, takes as definitive of the philosophical problem of scientific explanation..

4. Churchland's Connectionist Account. If this diagnosis is correct, then one should find similar problems at work for those theories of scientific explanation that keep the representational subsumption view in place but change the format of the representation. As a representative of psychologistic models of explanation more generally, consider Paul Churchland's (1989) parallel distributed processing (PDP) account of explanation. Churchland objects to Hempel's model (and, in fact, the entire logical empiricist enterprise) on the ground that human cognitive agents (such as scientists) do not in fact think with the structures of first order predicate logic. His revolutionary objective is to rebuild a model of science inspired by connectionist, or parallel distributed processing, theories of cognition rather than on 20th Century advances in logic.

On Churchland's view, understanding is prototype activation in a connectionist network:

> Explanatory understanding consists in the activation of a particular prototype vector in a well-trained network. It consists in the apprehension of the problematic case as an instance of a general type, a type for which the creature has a detailed and well-informed representation. (Churchland 1989, 210)

When we understand a phenomenon, we assimilate it to a prototype and thereby generate novel features of the phenomenon from a few input features. The prototype stores a wealth of theoretical information about a phenomenon. Understanding, accordingly, is a matter of recognizing that a given phenomenon fits a more general prototype. Scientific explanation involves the construction of prototypes (such as the HH model, presumably) that can be so applied.

The first thing to notice about Churchland's model of understanding is that he does not say how those instances of prototype-activation that constitute understanding are different from those that do not. Prototype activation vectors are used to describe many aspects of brain function. Stored patterns of activation across populations of neurons control balance, posture, and reaching; they produce and direct saccadic eye movements; and they regulate endocrine release and bodily fluid homeostasis. To put the point maximally bluntly: if the brain does it, it likely does it with activation vectors. So the idea that understanding involves the activation of prototype vectors tells us very little about the distinctive character of understanding.

To make this more concrete, consider the distinction between recognition and understanding. One can recognize Ike in a crowd without explaining anything about him. Suppose that one wants to understand why Ike is a bookie, or why Ike has only a junior high education. One cannot answer these questions by merely recognizing Ike. This is

because Ike's surface features (his gait, his hair line, his shape), that is, the kinds of things that will show up in the visual Ike-recognition vector, are in most cases not explanatorily relevant to his professional and educational status. To drive the point home, it would appear that Churchland's model does not have a principled means for distinguishing phenomenal models, which merely describe the phenomenon to be explained, from explanatory models, which explain why the *explanandum* phenomenon is as it is.

In the years since Churchland's suggestion, cognitive scientists have learned more about the cognitive mechanisms of causal understanding. Churchland could add further content to his account by building details about how human cognitive systems discern and represent the relevant ontic structures that constitute *bona fide* understanding. Though it is no trivial matter to formulate such a theory, there can be no doubt that such a theory could, in fact, be implemented in a connectionist network, whatever it is. However, notice that building a model of the cognitive capactieis that make bona fide understanding possible in creatures such as us requires one to say what the prototype vectors must be about in order to constitute bona fide understanding: that they are about causal structures, laws, statistical dependencies, mechanisms, or what have you. In other words, in order to say which specific cognitive capacities are relevant to our ability to understand the world, we must thrust our attention outward from representations to the ontic explanations that they must represent if they are to truly constitute understanding.

There is further reason to avoid equating scientific explanation with the abilities of individual cognitive agents. Some phenomena might be so complex that they overwhelm our limited (individual) cognitive systems. Perhaps a mechanism has so many

parts with so many interactions that it is impossible for a single person to fully understand. Perhaps scientists must rely on computer simulations, graphical representations, large compiled databases in order to build models that explain the complex phenomena in their domain. Perhaps human working memory is so limited that it cannot entertain all of the information explanatorily relevant to a given phenomenon (compare Rosenberg 1985; 1994). Mary Hegarty shows that even simple mechanisms overwhelm our processing capacities if they have over a handful of parts or if the interactions among them cannot be represented in two dimensions (Hegarty, Just, and Morrison 1988). For this reason, it seems inappropriate to model scientific explanation, which has no principled limit on its complexity, on the basis of individual human cognition, which is often quite limited. It would be wrong to say that phenomena produced by very complex mechanisms (i.e., those that outstrip our cognitive capacities) have no explanation. The explanations exist even if our brains cannot represent them.

Suppose, though, we accept Churchland's PDP model as an adequate account of the psychology of human understanding. Can this psychological account do double-duty as an account of the norms of scientific explanation? The inclusiveness of the PDP model (and the representational model in general) is again its primary drawback. The more permissive an account of explanatory representations, the less likely it is to fulfill the distinctions discussed above in Section 2. Churchland explicitly disavows interest in the norms of explanation (Churchland 1989, 198). However, the demands on a philosophical theory of explanation cannot be satisfied without thinking about norms for evaluating explanations. Consider Churchland's description of etiological causal prototypes:

> An etiological prototype depicts a typical temporal sequence of events, such as cooking of food upon exposure to heat, the deformation of a fragile object during impact with a tougher one, the escape of liquid form a tilted container, and so on. These sequences contain prototypical elements in a prototypical order, and they make possible our explanatory understanding of the temporally extended world. (Churchland 1989, 213)

But as discussed above, some temporal sequences are explanatory (if appropriately supplemented with the causal relations between the different events in the sequence), and some are not. An account of explanation should help one to distinguish the two. Churchland acknowledges this limitation: "Now just what intricacies constitute a genuine etiological prototype, and how the brain distinguishes between real causal processes and mere pseudoprocesses, are secondary matters I shall leave for a future occasion" (Churchland 1989, 214; 2005). Those who would develop a normative account of explanation, however, cannot avoid this question. The way to understand how brains distinguish causes from temporal sequences is to start by considering how causes differ from temporal sequences— that is, by examining the objective explanations in the world rather than the way that they are represented in the mind/brain. A similar point could be made about common cause structures and effect-to-cause explanations. That is, the model does not appear to have the resources to distinguish predictive models from explanatory models.

An equally fundamental problem arises when we consider the question of explanatory relevance. Grant that explanatory representations are prototypes and that explanation involves activating such prototypes. Different features of the phenomenon

are relevant for different explanatory purposes. Suppose that Ike is a member of the gang, the Sharks; he is single and 30 years old; he weighs 210 pounds; he has a junior high education; he is a bookie; he idolizes Johnny Ramone; and he plays guitar. To explain why he is a bookie, it would be relevant to note that he is a member of a gang and perhaps that he has a junior high education, but it would probably not be relevant to note that he weighs 210 pounds or that he plays guitar. To explain why he plays guitar, it might be relevant to note that he is a single, 30 year-old male who idolizes Johnny Ramone, but not (I suppose) that he is a bookie or that he has a junior high school education. All of these features are in the Ike prototype (which, if we know him well, contains innumerable other features of varying degrees of explanatory relevance to these phenomena). And all of these features are activated when we think of Ike. Yet only some of these features are relevant to explaining why he is a bookie, only some are relevant to explaining why he plays guitar, and few of the features in these two lists overlap.

What goes for Ike goes for the categories of science. Ion channels can be characterized along a number of dimensions: molecular weight, primary structure, voltage sensitivity, maximum conductance values, primary structure, and so on. Different features of a given type of ion channel are relevant for different explanatory purposes. An account of explanation that can be used to sort good explanations from bad should help to sort explanatorily relevant information from explanatorily irrelevant information. But the PDP account cannot be so used unless the activation-vector story is supplemented with an account of explanatory relevance. However, to supplement it, one will have to begin by assessing what explanatory relevance is, and this again thrusts our attention away from representation and out onto the ontic structures that good explanatory texts describe.

Hempel, who can be credited with initiating sustained philosophical discussion of the nature of scientific explanation, drew precisely the sharp line between explanation and understanding that I am here trying to make explicit:

> …man has long and persistently been concerned to achieve some understanding of the enormously diverse, often perplexing, and sometimes threatening occurrences in the world around him… Some of these explanatory ideas are based on anthropomorphic conceptions of the forces of nature, others invoke hidden powers or agents, still others refer to God's inscrutable plans or to fate.
>
> Accounts of this kind undeniably may give the questioner a sense of having attained some understanding; they may resolve his perplexity and in this sense 'answer' his question. But however satisfactory these answers may be psychologically, they are not adequate for the purposes of science, which, after all, is concerned to develop a conception of the world that has a clear, logical bearing on our experience and is capable of objective test. (1966, 47-8).

The point of this passage is to drive a wedge between the psychological mechanisms that give rise to the sense of intelligibility and understanding, on the one hand, and a properly philosophical theory of scientific explanation. The task is to develop an account of scientific explanation that makes sense of the scientific project of connecting our models to structures that can be discovered through experience and objective tests. In domains of science that concern themselves with the search for causes and mechanisms, this amounts to the idea that the norms of explanation fall out of a commitment by scientists to describe as accurately and completely as possible the relevant ontic structures in the

world. Viewed in this way, our theories of scientific explanation cannot carve off those ontic structures as if they were expendable in the search for a theory of explanation: the norms of explanation fall out of the scientific commitment to describe those ontic structures.

5. Idealization and the Ontic Conception. Let us now turn attention to the role of idealization in scientific explanation. As a matter of historical record, explanatory texts are often idealized in the sense that they make false assumptions about the system they represent in order to make the texts more compact and elegant. To make matters worse, such texts appear to function as they do in our scientific communication largely because they describe the relevant ontic structures incorrectly. If so, one might be tempted to conclude that it is inappropriate to emphasize the ontic mode of explanation; scientific explanation essentially involves divorcing one's thought from the relevant ontic structures and providing representations that make the messy phenomena intelligible and useful to creatures like us.

The undeniable fact that scientific models are typically idealized is clearly most problematic for accounts of explanation that demand that a scientific explanation must subsume a description of the phenomenon under a true general representation; that is, for the strongest versions of the representational subsumption view. Hempel, for example, requires as a criterion of adequacy on explanatory arguments that the premises of the argument be true. And for this reason, his model of explanation rather famously has difficulty accommodating the ubiquitous practice of idealization. Hempel was committed to a representational view, and to the idea that the representations in explanations have to

be true, so it was a real challenge for his view that explanatory models are (almost) always idealized.

Of course, the requirement that explanatory texts must be true is certainly reasonable. Even if one can subsume a description of the action potential under a model that posits the existence of black holes in the endoplasmic reticulum, and even if the model renders action potentials intelligible (that is, the model gives people the sense of understanding how action potentials are produced), such a model simply cannot explain the action potential. The reason is plain: there are no black holes in the endoplasmic reticulum. The ideal of scientific explanation cannot be wholly severed from the criterion of truth lest we lose any grip at all on the idea that it is a scientific explanation rather than an intelligible tale of some other sort. The goal of building an explanatory text is not to provide the illusion of understanding, but rather to provide *bona fide* understanding. Entirely false explanatory texts offer only the former. [5]

Idealized models, however, are of interest because they are not entirely false: they bring to light aspects of the system under investigation that are difficult to see unless one makes false assumptions. Things are easier if one assumes, for example, that the axon is cylindrical, that the concentration of ions is everywhere uniform, that the membrane obeys Ohm's law strictly. The explanatory text contains idealizing assumptions precisely because, in making such assumptions, one reveals aspects of the ontic structure of the system that would otherwise be occluded. The idealizing model thus has the capacity to

[5] The same point could be made in terms of empirical adequacy rather than truth, should that be preferred. Idealized theories, as I have described them, must be empirically inadequate in some respect; otherwise there would be no basis for the claim that they contain false assumptions.

inform us about the ontic explanations for phenomena even if model is not, strictly speaking, true.

Now, it would surely be a mistake to claim that a model has to be true to convey explanatory information. But conveying explanatory information about X and truly representing the explanation for X are not the same thing. Friends of the ontic conception should say that idealized models are useful for conveying true information about the explanation, but that they are not true representations of the explanation.

One benefit of clearly disambiguating the ontic mode from the communicative, representational, and cognitive modes of talking about explanation is that it allows us to divide labor on these matters. Terms like "true," "idealized," and "abstract" apply to representations or models. They do not apply to the ontic structures they represent (bracketing cases in which the ontic structures involved in the explanation are themselves representations). Once these are separated, the problem of idealization is clearly not a problem for philosophical theories of explanation; rather it is a problem for philosophical theories of description. The question at the heart of the problem of idealization is this: What is required for a given representation to convey information about the ontic structure of the world? This is an important question, but it is a question about reference, not a question about explanation. We only invite confusion if we fail to keep these questions distinct.

To say that a model is idealized is, *ipso facto*, to recognize a distinction between models that are true and models that are false. To say that a model is an idealization of an ontic explanation, after all, is to say that the model contains one or more false commitments about that ontic explanation. To very idea of an idealized model of an

explanation commits one, at least implicitly, to the existence of an ontic explanation against which the model can be evaluated. It is more sensible to say that idealized models convey explanatory information in virtue of making false assumptions that bring certain truths about the ontic explanation to light. If we say, in contrast, that false models explain, we are left scratching our heads about how a false model could be an explanation of anything at all. Our heads will itch, however, only if we are committed first and foremost to the idea that explanations are representations. But that is to get things backwards. The explanations are in the world. The scientist's task is to describe them. And they can use any number of representational tools to convey that explanatory information clearly and effectively. If we give up on the representational subsumption view as the heart of our philosophical theories of explanation, the problem of idealization then finds its proper home in semantics.

5. Conclusion. The central tasks for a philosophical theory of scientific explanation are a) to demarcate explanation from other kinds of scientific achievement, and b) to articulate the norms that distinguish adequate explanations from inadequate explanations. In this paper, I have argued that the term "explanation" is ambiguous, having at least four senses, and that one might construct a theory adequate to one of these senses without in the process constructing a theory that is adequate to the others. I have argued that the philosophical theory of explanation depends fundamentally on an ontic conception of explanation, that is, on a view about the kinds of structures in the world that count as legitimately explanatory. Appeal to such structures is required to distinguish how possibly from how actually explanations, phenomenal models from mechanistic

models, merely predictive models from explanatory models, sketches from complete-enough explanations, and relevant from irrelevant explanatory factors.

Just as representational views of explanation, on their own, cannot provide an account of the norms underlying a philosophical analysis of scientific explanation, an account that addresses those norms leaves work to be done by representational theories. Not all of the facts in an ontic explanation are salient in a given explanatory context, and for the purposes of communication, it is often necessary to abstract, idealize, and fudge to represent and communicate which ontic structures cause, constitute, or otherwise are responsible for such phenomena. Such topics are the proper province of psycholgistic theorizing about scientific explanation and work in the philosophy of reference. But these topics are separate from the classic philosophical topic of the nature of scientific explanation.

Bibliography.

Ayer, A.J. (1974). "What is a Law of Nature" in M. Curd and J.A. Cover, eds *Philosophy of Science: The Central Issues.* (New York: WW Norton Co.) 1998.

Bechtel, W., and A. Abrahamsen. (2005). "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Biological and Biomedical Sciences* 36,421-441.

Bogen, J. (2005). Regularities and causality; Generalizations and causal explanations. Forthcoming in *Studies in the History and Philosophy of Biology and the Biomedical Sciences*.

Bogen, J. (2008). "Causally Productive Activities." Studies *in History and Philosophy of Science Part A* 39 (1):112-123.

Churchland, P. M. (1989). *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science.* Cambridge, MA: MIT Press.

Coffa, J. A. (1974). ‘‘Hempel’s Ambiguity,’’ *Synthese*, 28: 141–63.

Craver C.F. (2006). “When mechanistic models explain.” *Synthese.* 153:355-376.

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

Craver, C.F. (2008). Physical law and mechanistic explanation in the Hodgkin and Huxley model of the action potential. *Philosophy of Science* 75:1022-33.

Dayan P & Abbott LF (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems.* Cambridge: MA. MIT Press.

Dray, W. (1957). *Laws and explanations in history.* Oxford: Oxford University Press.

Hegarty, M., Just, M. A. & Morrison, I. R. 1988. Mental models of mechanical systems: Individual differences in qualitative and quantitative reasoning. *Cognitive Psychology.* 20, 191-236.

Hempel, C. G. (1965). “Aspects of scientific explanation. In C.G. Hempel, ed., *Aspects of Scientific Explanation*. Free Press, New York. 331–496.

Hempel, C. G. (1966). *Philosophy of Natural Science.* Prentice Hall, NJ.

Hempel, C. G. and P. Oppenheim. (1948). Studies in the logic of explanation. *Philosophy of Science* 15:135–175.

Hille, B. 2001. *Ion channels of excitable membranes*. 3rd ed. Sinauer associates.

Hodgkin, A.L. and Huxley, A.F. (1952). “A quantitative description of membrane current and its application to conduction and excitation in nerve.” *Journal of Physiology* 117: 500-544.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science.* 48:507–531.

Machamer, Peter, Lindley Darden, and Carl F. Craver. (2000). "Thinking About Mechanisms." *Philosophy of Science* 67: 1-25.

Mitchell, Sandra D. (1997). "Pragmatic Laws." *Philosophy of Science* 64 (Proceedings):468-479.

Rosenberg, A. (1994). *Instrumental biology or the unity of science*. Chicago, IL: University of Chicago Press.

Rosenberg, A. (1985). *The structure of biological science*. Cambridge: Cambridge University Press.

Salmon, Wesley C. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, W. C. (1989). Four decades of scientific explanation. In Kitcher and Salmon (eds. *Scientific explanation, Minnesota Studies in the Philosophy of Science XVIII.* Minneapolis: University of Minnesota Press, pp. 3-219.

Thagard, Paul. (1999). *How scientists explain disease*. Princeton: Princeton University Press.

Woodward, J. (2003). *Making Things Happen*. New York: Oxford University Press.