D. The Explanatory Gap

35 Materialism and Qualia
The Explanatory Gap

Joseph Levine

In "Naming and Necessity"[1] and "Identity and Necessity,"[2] Kripke presents a version of the Cartesian argument against materialism. His argument involves two central claims: first, that all identity statements using rigid designators on both sides of the identity sign are, if true at all, true in all possible worlds where the terms refer; second, that psycho-physical identity statements are conceivably false, and therefore, by the first claim, actually false.

My purpose in this paper is to transform Kripke's argument from a metaphysical one into an epistemological one. My general point is this. Kripke relies upon a particular intuition regarding conscious experience to support his second claim. I find this intuition important, not least because of its stubborn resistance to philosophical dissolution. But I don't believe this intuition supports the meta-physical thesis Kripke defends—namely, that psycho-physical identity statements must be false. Rather, I think it supports a closely related epistemological thesis—namely, that psycho-physical identity statements leave a significant *explanatory gap,* and, as a corollary, that we don't have any way of determining exactly which psycho-physical identity statements are true.[3] One cannot conclude from my version of the argument that materialism is false, which makes my version a weaker attack than Kripke's. Nevertheless, it does, if correct, constitute a problem for materialism, and one that I think better captures the uneasiness many philosophers feel regarding that doctrine.

I will present this epistemological argument by starting with Kripke's own argument and extracting the underlying intuition. For brevity's sake, I am going to assume knowledge of Kripke's general position concerning necessity and the theory of reference, and concentrate only on the argument against materialism. To begin with, let us assume that we are dealing with a physicalist type-identity theory. That is, our materialist is committed to statements like:

(1) Pain is the firing of C-fibers.

On Kripke's general theory, if (1) is true at all it is necessarily true. The same of course, is the case with the following statement:

(2) Heat is the motion of molecules.

That is, if (2) is true at all it is necessarily true. So far so good.

The problem arises when we note that, with both (1) and (2), there is a felt contingency about them. That is, it seems conceivable that they be false. If they are necessarily true, however, that means there is no possible world in which they are false. Thus, imagining heat without the motion of molecules, or pain without the firing of C-fibers, must be to imagine a logically impossible world. Yet these suppositions *seem* coherent enough. Kripke responds that the felt contingency of (2) can be satisfactorily explained away, but that this can't be done for (1). Thus, there is an important difference between psycho-physical identities and other theoretical identities, and this difference makes belief in the former implausible.

The difference between the two cases is this. When it seems plausible that (2) is contingent, one can become disabused of this notion by noting that instead of imagining *heat* without the motion of molecules, one is really imagining there being some phenomenon that affects our senses the way heat in fact does, but is not the motion of molecules. The truly contingent statement is not (2) but

From *Pacific Philosophical Quarterly* 64:354–61, 1983. Reprinted with permission of author and publisher. Addendum excerpted from "On Leaving Out What It's Like," in M. Davies & G. Humphreys, eds., *Consciousness* (Blackwell, 1993), with permission of author and publisher.

(2′) The phenomenon we experience through the sensations of warmth and cold, which is responsible for the expansion and contraction of mercury in thermometers, which causes some gases to rise and others to sink, etc., is the motion of molecules.

However, this sort of explanation will not work for (1). When we imagine a possible world in which a phenomenon is experienced as pain but we have no C-fibers, that is a possible world in which there *is* pain without there being any C-fibers. This is so, argues Kripke, for the simple reason that the experience of pain, the sensation of pain, counts as pain itself. We cannot make the distinction here, as we can with heat, between the way it appears to us and the phenomenon itself. Thus, we have no good account of our intuition that (1) is contingent, unless we give up the truth of (1) altogether.

Now, there are several responses available to the materialist. First of all, the most popular materialist view nowadays is functionalism, which is not committed to even the contingent truth of statements like (1). Rather than identifying types of mental states with types of physical states, functionalists identify the former with types of functional, or what Boyd calls "configurational" states.[4] Functional states are more abstract than physical states, and are capable of realization in a wide variety of physical constitutions. In terms of the computer metaphor, which is behind many functionalist views, our mentality is a matter of the way we are "programmed," our "software," whereas our physiology is a matter of our "hardware." On this view, the intuition that pain could exist without C-fibers is explained in terms of the multiple realizability of mental states. This particular dilemma, then, doesn't appear to arise for functionalist materialists.

However, this reply won't work. First of all, a Kripke-style argument can be mounted against functionalist identity statements as well. Ned Block, in "Troubles with Functionalism,"[5] actually makes the argument. He asks us to imagine any complete functionalist description of pain (embedded, of course, in a relatively complete functionalist psychological theory). Though we have no idea as yet exactly what this description would be, insofar as it is a *functionalist* description, we know roughly what form it would take. Call this functionalist description "F." Then functionalism entails the following statement:

(3) To be in pain is to be in state F.

Again, on Kripke's theory of reference, (3) is necessarily true if true at all. Again, it seems imaginable that in some possible world (perhaps even in the actual world) (3) is false. Block attempts to persuade us of this by describing a situation where some object is in F but it is doubtful that it is in pain. For instance, suppose F were satisfied by the entire nation of China—which, given the nature of functional descriptions, is logically possible. Note that all the argument requires is that it should be *possible* that the entire nation of China, while realizing F, not be in pain. This certainly does seem possible.

Furthermore, some adherents of functionalism have moved back toward physicalist reductionism for qualia, largely in response to considerations like those put forward by Block. The idea is this. What Block's example seems to indicate is that functional descriptions are just *too* abstract to capture the essential features of qualitative sensory experiences. The so-called "inverted spectrum" argument—which involves the hypothesis that two people could share functional descriptions yet experience different visual qualia when viewing the same object—also points up the excessive abstractness of functional descriptions. Now one way some functionalists propose to deal with this problem is to return to a physicalist type-identity theory for sensory qualia, or at least for particular kinds of sensory qualia.[6] The gist of the latter proposal is this. While it's sufficient for being conscious (for having qualia at all) that an entity realize the appropriate functional description, the particular way a qualitative state is experienced is determined by the nature of the physical realization. So if, while looking at a ripe McIntosh apple, I experience the visual quality normally associated with looking at ripe McIntosh apples, and my inverted friend experiences the quality normally associated with looking at ripe cucumbers, this has to do with the difference in our physical realizations of the same functional state. Obviously, if we adopt this position Kripke's original argument applies.

So far, then, we see that the move to functionalism doesn't provide materialists with a way to avoid the dilemma Kripke poses: either bite the bullet and deny that (1), or (3), is contingent, or give up materialism. Well, what about biting the bullet? Why not just say that, intuition notwithstanding, statements like (1) and (3) are not contingent? In fact, Kripke himself, by emphasiz-

ing the gulf between epistemological possibility and metaphysical possibility, might even seem to give the materialist the ammunition she needs to attack the legitimacy of the appeal to this intuition. For what seems intuitively to be the case is, if anything, merely an epistemological matter. Since epistemological possibility is not sufficient for metaphysical possibility, the fact that what is intuitively contingent turns out to be metaphysically necessary should not bother us terribly. It's to be expected.

In the end, of course, one can just stand pat and say that. This is why I don't think Kripke's argument is entirely successful. However, I do think the intuitive resistance to materialism brought out by Kripke (and Block) should not be shrugged off as *merely* a matter of epistemology. Though clearly an epistemological matter, I think this intuitive resistance to materialism should bother us a lot. But before I can defend this claim, the intuition in question requires some clarification.

First of all, let's return to our list of statements. What I want to do is look more closely at the difference between statement (2) on the one hand, and statements (1) and (3) on the other. One difference between them, already noted, was the fact that the felt contingency of (2) could be explained away while the felt contingency of the others could not. But I want to focus on another difference, one which I think underlies the first one. Statement (2), I want to say, expresses an identity that is *fully explanatory,* with nothing crucial left out. On the other hand, statements (1) and (3) do seem to leave something crucial unexplained, there is a "gap" in the explanatory import of these statements. It is this explanatory gap, I claim, which is responsible for their vulnerability to Kripke-type objections. Let me explain what I mean by an "explanatory gap."

What is explanatory about (2)? (2) states that heat is the motion of molecules. The explanatory force of this statement is captured in statements like (2′) above. (2′) tells us by what mechanism the causal functions we associate with heat are effected. It is explanatory in the sense that our knowledge of chemistry and physics makes intelligible how it is that something like the motion of molecules could play the causal role we associate with heat. Furthermore, antecedent to our discovery of the essential nature of heat, its causal role, captured in statements like (2′), exhausts our notion of it. Once we understand how this causal role is carried out there is nothing more we need to understand.

Now, what is the situation with (1)? What is explained by learning that pain is the firing of C-fibers? Well, one might say that in fact quite a bit is explained. If we believe that part of the concept expressed by the term "pain" is that of a state which plays a certain causal role in our interaction with the environment (e.g. it warns us of damage, it causes us to attempt to avoid situations we believe will result in it, etc.), (2) explains the mechanisms underlying the performance of these functions. So, for instance, if penetration of the skin by a sharp metallic object excites certain nerve endings, which in turn excite the C-fibers, which then causes various avoidance mechanisms to go into effect, the causal role of pain has been explained.

Of course, the above is precisely the functionalist story. Obviously, there is something right about it. Indeed, we do feel that the causal role of pain is crucial to our concept of it, and that discovering the physical mechanism by which this causal role is effected explains an important facet of what there is to be explained about pain. However, there is more to our concept of pain than its causal role, there is its qualitative character, how it feels; and what is left unexplained by the discovery of C-fiber firing is *why pain should feel the way it does!* For there seems to be nothing about C-fiber firing which makes it naturally "fit" the phenomenal properties of pain, any more than it would fit some other set of phenomenal properties. Unlike its functional role, the identification of the qualitative side of pain with C-fiber firing (or some property of C-fiber firing) leaves the connection between it and what we identify it with completely mysterious. One might say, it makes the way pain feels into merely a brute fact.

Perhaps my point is easier to see with the example above involving vision. Let's consider again what it is to see green and red. The physical story involves talk about the various wavelengths detectable by the retina, and the receptors and processors that discriminate among them. Let's call the physical story for seeing red "R" and the physical story for seeing green "G." My claim is this. When we consider the qualitative character of our visual experiences when looking at ripe McIntosh apples, as opposed to looking at ripe cucumbers, the difference is not explained by appeal to G and R. For R doesn't really explain why I have the one kind of qualitative experience—the kind I have when looking at McIntosh apples—and not the other. As evidence for this, note that it seems just as easy

to imagine G as it is to imagine R underlying the qualitative experience that is in fact associated with R. The reverse, of course, also seems quite imaginable.

It should be clear from what's been said that it doesn't help if we actually identify qualia with their functional roles. First of all, as I mentioned above, some functionalists resist this and prefer to adopt some form of type-physicalism for qualia. So when seeking the essence of how it feels to be in a certain functional state, they claim we must look to the essence of the physical realization. Secondly, even if we don't take this route, it still seems that we can ask why the kind of state that performs the function performed by pain, whatever its physical basis, should *feel* the way pain does. The analogous question regarding heat doesn't feel compelling. If someone asks why the motion of molecules plays the physical role it does, one can properly reply that an understanding of chemistry and physics is all that is needed to answer that question. If one objects that the phenomenal properties we associate with heat are not explained by identifying it with the motion of molecules, since being the motion of molecules seems compatible with all sorts of phenomenal properties, this just reduces to the problem under discussion. For it is precisely phenomenal properties—how it is for us to be in certain mental (including perceptual) states—which seem to resist physical (including functional) explanations.

Of course, the claim that (1) and (3) leave an explanatory gap in a way that (2) doesn't cannot be made more precise than the notion of explanation itself. Obviously, the D-N model of explanation is not sufficient for my purposes, since (1) and (3) presumably support counterfactuals and could be used, along with other premises, to deduce all sorts of particular facts.[7] What we need is an account of what it is for a phenomenon to be made *intelligible,* along with rules which determine when the demand for further intelligibility is inappropriate. For instance, I presume that the laws of gravity explain, in the sense at issue here, the phenomena of falling bodies. There doesn't seem to be anything "left out." Yet I am told that the value of *G,* the gravitational constant, is not derived from any basic laws. It is a given, a primitive, brute fact about the universe. Does this leave us with a feeling that something which ought to be explained is not? Or do we expect that some facts of nature should appear arbitrary in this way? I am inclined to take the latter attitude with respect to *G.* So, one may ask, why does the connection between what it's like to be in a particular functional (or physical) state and the state itself demand explanation, to be made intelligible?

Without a theoretical account of the notion of intelligibility I have in mind, I can't provide a really adequate answer to this question. Yet I think there are ways to at least indicate why it is reasonable to seek such an explanation. First of all, the phenomenon of consciousness arises on the macroscopic level. That is, it is only highly organized physical systems which exhibit mentality. This is of course what one would expect if mentality were a matter of functional organization. Now, it just seems odd that primitive facts of the sort apparently presented by statements like (1) and (3) should arise at this level of organization. Materialism, as I understand it, implies explanatory reductionism of at least this minimal sort: that for every phenomenon not describable in terms of the fundamental physical magnitudes (whatever they turn out to be), there is a mechanism that is describable in terms of the fundamental physical magnitudes such that occurrences of the former are intelligible in terms of occurrences of the latter. While this minimal reductionism does not imply anything about the reducibility of theories like psychology to physics, it does imply that brute facts—of the sort exemplified by the value of *G*—will not arise in the domain of theories like psychology.

Furthermore, to return to my original point, the claim that statements (1) and (3) leave an explanatory gap accounts for their apparent contingency, and, more importantly, for the failure to explain away their apparent contingency in the standard way. After all, why is it that we can account for the apparent contingency of (2) in a theoretically and intuitively satisfactory manner, but not for that of (1) and (3)? Even if one believes that we don't have to take this intuitive resistance seriously, it is still legitimate to ask why the problem arises in these particular cases. As I claimed above, I think the difference in this regard between (2) on the one hand, and (1) and (3) on the other, is accounted for by the explanatory gap left by the latter as opposed to the former. Since this is the crucial connection between Kripke's argument and mine, let me belabor this point for a bit.

The idea is this. If there is nothing we can determine about C-fiber firing that explains why having one's C-fibers fire has the qualitative character that it does—or, to put it another way,

if what it's particularly like to have one's C-fibers fire is not explained, or made intelligible, by understanding the physical or functional properties of C-fiber firings—it immediately becomes imaginable that there be C-fiber firings without the feeling of pain, and vice versa. We don't have the corresponding intuition in the case of heat and the motion of molecules—once we get clear about the right way to characterize what we imagine—because whatever there is to explain about heat is explained by its being the motion of molecules. So, how could it be anything else?

The point I am trying to make was captured by Locke[8] in his discussion of the relation between primary and secondary qualities. He states that the simple ideas which we experience in response to impingements from the external world bear no intelligible relation to the corpuscular processes underlying impingement and response. Rather, the two sets of phenomena—corpuscular processes and simple ideas—are stuck together in an arbitrary manner. The simple ideas go with their respective corpuscular configurations because God chose to so attach them. He could have chosen to do it differently. Now, so long as the two states of affairs seem arbitrarily stuck together in this way, imagination will pry them apart. Thus it is the non-intelligibility of the connection between the feeling of pain and its physical correlate that underlies the apparent contingency of that connection.

Another way to support my contention that psycho-physical (or psycho-functional) identity statements leave an explanatory gap will also serve to establish the corollary I mentioned at the beginning of this paper; namely, that even if some psycho-physical identity statements are true, we can't determine exactly which ones are true. The two claims, that there is an explanatory gap and that such identities are, in a sense, unknowable, are interdependent and mutually supporting. First I will show why there is a significant problem about our ever coming to know that statements like (1) are true, then I will show how this is connected to the problem of the explanatory gap.

So suppose, as a matter of fact, that having the feeling of pain is identical with being in a particular kind of physical state. Well, which physical state? Suppose we believed it to be the firing of C-fibers because that was the state we found to be correlated with the feeling of pain in ourselves. Now imagine we come across alien life which gives every behavioral and functional sign of sharing our qualitative states. Do they have the feeling of pain we have? Well, if we believed that to have that feeling is to have one's C-fibers fire, and if the aliens don't have firing C-fibers, then we must suppose that they can't have this feeling. But the problem is, even if it is true that creatures with physical constitutions radically different from ours do not share our qualitative states, how do we determine what measure of physical similarity/dissimilarity to use? That is, the fact that the feeling of pain is a kind of physical state, if it is, doesn't itself tell us how thickly or thinly to slice our physical kinds when determining which physical state it is identical to. For all we know, pain is identical to the disjunctive state, the firing of C-fibers *or* the opening of D-valves (the latter disjunct realizing pain [say] in creatures with a hydraulic nervous system).[9]

This objection may seem like the standard argument for functionalism. However, I am actually making a quite different argument. First of all, the same objection can be made against various forms of functionalist identity statements. That is, if we believe that to have the feeling of pain is to be in some functional state, what measure of functional similarity/dissimilarity do we use in judging whether or not some alien creature shares our qualitative states? Now, the more inclusive we make this measure, the more pressure we feel about questions of inverted qualia, and therefore the more reason we have to adopt a physicalist-reductionist position concerning particular kinds of qualia. This just brings us back where we started. That is, if having a radically different physical constitution is sufficient for having different qualia, there must be some fact of the matter about *how* different the physical constitution must be. But what possible evidence could tell between the hypothesis that the qualitative character of our pain is a matter of having firing C-fibers, and the hypothesis that it is a matter of having either firing C-fibers or opening D-valves?[10]

Now, if there were some intrinsic connection discernible between having one's C-fibers firing (or being in functional state F) and what it's like to be in pain, by which I mean that experiencing the latter was intelligible in terms of the properties of the former, then we could derive our measure of similarity from the nature of the explanation. Whatever properties of the firing of C-fibers (or being in state F) that explained the feel of pain would determine the properties a kind of physical (or functional) state had to have

our qualitative states. [illegible] f pain we have? Well, if we [illegible] e that feeling is to have [illegible] if the aliens don't have [illegible] must suppose that they [illegible] But the problem is, even if [illegible] s with physical constitution [illegible] from ours do not share [illegible] how do we determine what [illegible] cal similarity/dissimilarity [illegible] ict that the feeling of pain is a [illegible] tate, if it is, doesn't itself [illegible] thinly to slice our physical [illegible] ining which physical states [illegible] all we know, pain is identical [illegible] state, the firing of C-fibers [illegible] alves (the latter disputed [illegible] n creatures with a hydraulic [illegible]

ıay seem like the standard [illegible] nalism. However, I am not [illegible] e different argument. First of [illegible] ion can be made against [illegible] tionalist identity statements [illegible] ve that to have the feeling of [illegible] ome functional state, what [illegible] onal similarity/dissimilarity [illegible] ıg whether or not some alien [illegible] qualitative states? Now, to [illegible] make this measure, the most [illegible] ıbout questions of intensity [illegible] e the more reason we have to [illegible] t-reductionist position and [illegible] kinds of qualia. This was [illegible] re we started. That is, if two [illegible] rent physical constitutions [illegible] ; different qualia, there must [illegible] matter about *how* different [illegible] ıtion must be. But what pay [illegible] l tell between the hypothesis [illegible] character of our pain is a [illegible] ng C-fibers, and the hypothesis [illegible] tter of having either firing [illegible] D-valves?[10]

e some intrinsic connection [illegible] having one's C-fibers firing [illegible] ıal state F) and what it's like [illegible] ch I mean that experience [illegible] gible in terms of the physical [illegible] then we could derive [illegible] y from the nature of the [illegible] properties of the firing of [illegible] ı state F) that explained the [illegible] determine the properties of [illegible] unctional) state had to have [illegible]

in order to count as feeling like our pain. But without this explanatory gap filled in, facts about the kind or the existence of phenomenal experiences of pain in creatures physically (or functionally) different from us become impossible to determine. This, in turn, entails that the truth or falsity of (1), while perhaps metaphysically factual, is nevertheless epistemologically inaccessible. This seems to be a very undesirable consequence of materialism.

There is only one way in the end that I can see to escape this dilemma and remain a materialist. One must either deny, or dissolve, the intuition which lies at the foundation of the argument. This would involve, I believe, taking more of an eliminationist line with respect to qualia than many materialist philosophers are prepared to take. As I said earlier, this kind of intuition about our qualitative experience seems surprisingly resistant to philosophical attempts to eliminate it. As long as it remains, the mind/body problem will remain.[11]

ADDENDUM: FROM "ON LEAVING OUT WHAT IT'S LIKE"

This difference between the two cases reflects an important epistemological difference between the purported reductions of water to H_2O and pain to the firing of C-fibres: namely, that the chemical theory of water explains what needs to be explained, whereas a physicalist theory of qualia still 'leaves something out.' It is because the qualitative character itself is left *unexplained* by the physicalist or functionalist theory that it remains conceivable that a creature should occupy the relevant physical or functional state and yet not experience qualitative character.

The basic idea is that a reduction should explain what is reduced, and the way we tell whether this has been accomplished is to see whether the phenomenon to be reduced is epistemologically necessitated by the reducing phenomenon, i.e. whether we can see why, given the facts cited in the reduction, things must be the way they seem on the surface. I claim that we have this with the chemical theory of water but not with a physical or functional theory of qualia. The robustness of the absent and inverted qualia intuitions is testimony to this lack of explanatory import.

Let me make the contrast between the reduction of water to H_2O and a physico-functional reduction of qualia more vivid. What is explained by the theory that water is H_2O? Well, as an instance of something that's explained by the reduction of water to H_2O, let's take its boiling point at sea level. The story goes something like this. Molecules of H_2O move about at various speeds. Some fast-moving molecules that happen to be near the surface of the liquid have sufficient kinetic energy to escape the intermolecular attractive forces that keep the liquid intact. These molecules enter the atmosphere. That's evaporation. The precise value of the intermolecular attractive forces of H_2O molecules determines the vapour pressure of liquid masses of H_2O, the pressure exerted by molecules attempting to escape into saturated air. As the average kinetic energy of the molecules increases, so does the vapour pressure. When the vapour pressure reaches the point where it is equal to atmospheric pressure, large bubbles form within the liquid and burst forth at the liquid's surface. The water boils.

I claim that given a sufficiently rich elaboration of the story above, it is inconceivable that H_2O should not boil at 212°F at sea level (assuming, again, that we keep the rest of the chemical world constant). But now contrast this situation with a physical or functional reduction of some conscious sensory state. No matter how rich the information processing or the neurophysiological story gets, it still seems quite coherent to imagine that all that should be going on without there being anything it's like to undergo the states in question. Yet, if the physical or functional story really explained the qualitative character, it would not be so clearly imaginable that the qualia should be missing. For, we would say to ourselves something like the following:

> Suppose creature *X* satisfies functional (or physical) description *F*. I understand—from my functional (or physical) theory of consciousness—what it is about instantiating *F* that is responsible for its being a conscious experience. So how could *X* occupy a state with those very features and yet *not* be having a conscious experience?

The Conceptual Basis of the Explanatory Gap

I have argued that there is an important difference between the identification of water with

H_2O, on the one hand, and the identification of qualitative character with a physico-functional property on the other. In the former case the identification affords a deeper understanding of what water is by explaining its behaviour. Whereas, in the case of qualia, the subjective character of qualitative experience is left unexplained, and therefore we are left with an incomplete understanding of that experience. The basis of my argument for the existence of this explanatory gap was the conceivability of a creature's instantiating the physico-functional property in question while not undergoing an experience with the qualitative character in question, or any qualitative character at all.

In order fully to appreciate the nature and scope of the problem, however, it is necessary to explore in more detail the basis of the explanatory adequacy of theoretical reductions such as that of water to H_2O, as well as the difference between these cases and the case of qualitative character. I can only begin that project here, with the following admittedly sketchy account. We will see that an adequate account must confront deep problems in the theory of conceptual content, thus drawing a connection between the issue of intentionality and the issue of consciousness.

Explanation and Reduction

To begin with, it seems clear that theoretical reduction is justified principally on the basis of its explanatory power. For instance, what justifies the claim that water is H_2O anyway? Well, we might say that we find a preponderance of H_2O molecules in our lakes and oceans, but of course that can't be the whole story. First of all, given all the impurities in most samples of water, this may not be true. Second, if we found that everything in the world had a lot of H_2O in it—suppose H_2O were as ubiquitous as protons—we wouldn't identify *water* with H_2O. Rather, we justify the claim that water is H_2O by tracing the causal responsibility for, and the explicability of, the various superficial properties by which we identify water—its liquidity at room temperature, its freezing and boiling points etc.—to H_2O.

But suppose someone pressed further, asking why being causally responsible for, this particular syndrome of superficial properties should be so crucial.[12] Well, we would say, *what else* could it take to count as water? But the source of this 'what else' is obscure. In fact, I think we have to recognize an a priori element in our justification. That is, what justifies us in basing the identification of water with H_2O on the causal responsibility of H_2O for the typical behaviour of water is the fact that our very concept of water is of a substance that plays such-and-such a causal role. To adopt Kripke's terminology, we might say that our pretheoretic concept of water is characterizable in terms of a 'reference fixing' description that roughly carves out a causal role. When we find the structure that in this world occupies that role, then we have the referent of our concept.

But now how is it that we get an explanation of these superficial properties from the chemical theory? Remember, explanation is supposed to involve a deductive relation between explanans and explanandum. The problem is that chemical theory and folk theory don't have an identical vocabulary, so somewhere one is going to have to introduce bridge principles. For instance, suppose I want to explain why water boils, or freezes, at the temperatures it does. In order to get an explanation of these facts, we need a definition of 'boiling' and 'freezing' that brings these terms into the proprietary vocabularies of the theories appealed to in the explanation.

Well, the obvious way to obtain the requisite bridge principles is to provide theoretical reductions of these properties as well.[13] To take another example, we say that one of water's superficial properties is that it is colourless. But being colourless is not a chemical property, so before we can explain why water is colourless in terms of the molecular structure of water and the way that such structures interact with light waves, we need to reduce colourlessness to a property like having a particular spectral reflectance function. Of course, the justification for this reduction will, like the reduction of water to H_2O, have to be justified on grounds of explanatory enrichment as well. That is, there are certain central phenomena we associate with colour, by means of which we pick it out, such that explaining those phenomena is a principal criterion for our acceptance of a theoretical reduction of colour.

The picture of theoretical reduction and explanation that emerges is of roughly the following form. Our concepts of substances and properties like water and liquidity can be thought of as representations of nodes in a network of causal relations, each node itself capable of further reduction to yet another network, until we get down to the fundamental causal determinants of nature. We get bottom-up necessity,

and thereby explanatory force, from the identification of the macroproperties with the microproperties because the network of causal relations constitutive of the micro level realizes the network of causal relations constitutive of the macro level. Any concept that can be analysed in this way will yield to explanatory reduction.

Notice that on this view explanatory reduction is, in a way, a two-stage process. Stage 1 involves the (relatively? quasi?) a priori process of working the concept of the property to be reduced 'into shape' for reduction by identifying the causal role for which we are seeking the underlying mechanisms. Stage 2 involves the empirical work of discovering just what those underlying mechanisms are.[14] . . .

NOTES

1. Saul Kripke, "Naming and Necessity," reprinted in *Semantics of Natural Language,* second edition, edited by Donald Davidson and Gilbert Harman, D. Reidel Publishing Co., 1972.
2. Saul Kripke, "Identity and Necessity," reprinted in *Naming, Necessity, and Natural Kinds,* edited by Stephen Schwartz, Cornell U. Press, 1977.
3. My argument in this paper is influenced by Thomas Nagel's in his paper "What Is It Like to Be a Bat?" (reprinted in *Readings in the Philosophy of Psychology,* volume 1, edited by Ned Block, Harvard U. Press, 1980), as readers who are familiar with Nagel's paper will notice as it develops.
4. Richard Boyd, "Materialism without Reductionism," reprinted in *Readings in the Philosophy of Psychology,* volume 1.
5. Ned Block, "Troubles with Functionalism," reprinted in *Readings in the Philosophy of Psychology,* volume 1.
6. Cf. Sydney Shoemaker, "The Inverted Spectrum," *The Journal of Philosophy,* volume LXXIX, no. 7, July 1982.
7. To elaborate a bit, on the D-N model of explanation, a particular event *e* is explained when it is shown to be deducible from general laws together with whatever description of the particular situation is relevant. Statements (1) and (3) could obviously be employed as premises in a deduction concerning (say) someone's psychological state. Cf. Carl Hempel, "Aspects of Scientific Explanation," reprinted in Hempel, *Aspects of Scientific Explanation,* Free Press, 1968.
8. Cf. Locke, *An Essay Concerning Human Understanding,* edited by J. Yolton, Everyman's Library, 1971 (originally published 1690); Bk. II, Ch. VIII, sec. 13, and Bk. IV, Ch. III, secs. 12 and 13.
9. This point is similar to an argument of Putnam's in the chapter of *Reason, Truth, and History* (Cambridge U. Press, 1981) entitled "Mind and Body." Putnam uses the argument to serve a different purpose from mine, however. The example of the hydraulic nervous system is from David Lewis, "Mad Pain and Martian Pain," reprinted in *Readings in the Philosophy of Psychology,* volume 1.
10. Shoemaker, in "The Inverted Spectrum," op. cit. explicitly tries to deal with this problem. He proposes a fairly complicated principle according to which disjunctive states like the one mentioned in the text do not qualify for identification with (or realization of) qualitative states. I cannot discuss his principle in detail here. However, the main idea is that we look to the causal role of a quale for its individuation conditions. That is, if the causal effects of pain in human beings are explained by their C-fiber firings *alone,* then the state of having one's C-fibers fire *or* having one's D-valves open is not a legitimate candidate for the physical realization of pain. Viewed from the standpoint of my argument in this paper, Shoemaker's principle begs the very question at issue; namely, whether the qualitative character of pain is explained by its causal role. For if it isn't, there is no reason to presume that the identity conditions of the physical state causally responsible for pain's functional role would determine the presence or absence of a particular kind of qualitative character. So long as the nature of that qualitative character is not explained by anything peculiar to any particular physical realization of pain, we have no way of knowing whether or not a different physical realization of pain, in a different creature, is associated with the same qualitative character.
11. An earlier version of this paper, under the title "Qualia, Materialism, and the Explanatory Gap," was delivered at the APA Eastern Division meetings, 1982. I would like to thank Carolyn McMullen for her comments on that occasion. I would also like to thank Louise Antony, Hilary Putnam, and Susan Wolf for their helpful comments on even earlier versions.
12. Of course, it's possible to imagine situations in which we would accept a theory of water that nevertheless left many of its superficial properties unexplained. However, unless the theory explained at least some of these properties, it would be hard to say why we consider this a theory of *water.*
13. In some cases, for instance with properties such as liquidity and mass, it might be better to think of their theoretical articulations in physical and chemical theory more as a matter of incorporating and refining folk theoretic concepts than as a matter of reducing them. But this is not an idea I can pursue here.
14. To a certain extent my argument here is similar to Alan Sidelle's defence of conventionalism in *Necessity, Essence, and Individuation* (Cornell University Press, 1989) though I don't believe our positions coincide completely.