34 On Leaving Out What It's Like

Joseph Levine

Among the reasons for doubting the adequacy of physicalist theories of the mind is the charge that such theories must "leave out" the qualitative, conscious side of mental life. One problem with evaluating this objection to physicalism is that it is not clear just what physicalist theories are being charged with. What is it for a theory to "leave out" a phenomenon? My project in this chapter is threefold: First, I want to clarify the anti-physicalist charge of "leaving out" qualia, distinguishing between a metaphysical and an epistemological reading of the objection. Second, I will argue that standard anti-physicalist conceivability arguments fail to show that physicalist theories "leave out" qualia in the metaphysical sense. But, third, I will also argue that these conceivability arguments do serve to establish that physicalist theories "leave out" qualia in the epistemological sense, because they reveal our inability to explain qualitative character in terms of the physical properties of sensory states. The existence of this "explanatory gap" constitutes a deep inadequacy in physicalist theories of the mind.[1]

The Metaphysical Reading

To begin, let us focus on the metaphysical reading of the phrase "leave out." In this sense, to say that a theory leaves out a certain phenomenon is to say that there are objects, events or properties to which the descriptive apparatus of the theory cannot refer. For instance, on Descartes's view, since the mind is composed of a non-physical, unextended substance, there is no way to use the predicates that apply to extended objects to refer to the mind. Property dualist views are similar in this respect. For a property dualist, there is no way of constructing descriptions using physical predicates[2] that apply to mental properties.

At least since Descartes, anti-physicalist arguments have taken roughly the following form. It is alleged that certain situations are imaginable, conceivable etc., and then a metaphysical conclusion is drawn. So, Descartes claims that from the fact that he can coherently conceive of the situation in which his body does not exist—for example, he may be deceived by an evil demon—and from the fact that he cannot conceive of the situation in which his mind does not exist (i.e., consistent with his having his current experiences), it follows that his body and his mind are not identical.

A look at the current state of the debate shows that anti-physicalist arguments have not advanced significantly beyond Descartes's. In particular, I want to focus on the two most prominent contemporary anti-physicalist arguments, those of Saul Kripke (1980) and Frank Jackson (1982).

Kripke's Argument

Kripke argues that there is an important asymmetry between purported mental-physical identity statements and those that derive from other scientific reductions. In both cases, if the identity statements are true, they are necessarily true. Also, in both cases, the identity statements involved appear contingent.[3] The asymmetry arises when we attempt to explain away their apparent contingency. Whereas the apparent contingency of other scientific identity statements can be explained away adequately, this cannot be done for mental-physical identity statements.

Suppose we compare a standard scientific identity statement like (1) below to a mental-physical identity statement like (2) below:

1. Water = H_2O
2. Pain = the firing of C-fibers

Since neither statement is known *a priori*, they are both imaginably false. Yet, if they are true, they

are necessarily true—they are not even possibly false. How do we reconcile the apparent contingency with the actual necessity? According to Kripke, this is easy to do in the case of (1). When we think we are imagining a situation in which water is not H_2O, in fact we are imagining a situation in which some substance which behaves superficially like water—but is not *water*—is not H_2O. On the other hand, a similar account will not work to explain the apparent contingency of (2), for to imagine a situation in which one is experiencing a state superficially like pain *just is* to imagine a situation in which one is experiencing pain. Conscious mental states are unlike external objects in that the standard distinction between how they appear and how they really are does not apply.

Many responses to Kripke's argument have appeared over the years. Early on it was pointed out that materialism does not entail the sort of type-type reductionism of the mental to the physical that is manifested in statements like (2). Rather, mental states are higher-order functional states, which can be realized, at least in principle, in a wide variety of physical systems. Hence, it is quite consistent with materialism that it is possible for one to experience pain and yet have no C-fibers whatever to fire.

However, functionalism itself has come under attack from Cartesian-style objections, particularly the inverted and absent qualia hypotheses.[4] The essence of these objections is that it seems perfectly imaginable that there could be creatures functionally alike who nevertheless differed in the qualitative character of their experiences; or, even worse, that there could be a creature functionally like ourselves who had no qualitative experiences at all. One line of response to either or both of these objections is to retreat to a physiological reductionist view with respect to qualia. That is, instead of identifying qualia with functional states, we identify them with the neurophysiological states that play the relevant functional roles in human beings, which would explain the possibility of both inverted and absent qualia.[5] Of course, this just brings us back to where we started.

I favor another strategy in response to Kripke's argument. Suppose he's right that we can coherently imagine feeling pain without having C-fibers firing. What's more, suppose he's right that this coherently imagined scenario cannot be explained away in the manner in which we explain away imagining that water is not H_2O. Still, what is imaginable is an *epistemological* matter, and therefore what imagining pain without C-fibers does is establish the *epistemological* possibility that pain is not identical with the firing of C-fibers. It takes another argument to get from the *epistemological* possibility that pain is not the firing of C-fibers to the metaphysical possibility, which is what you need to show that pain isn't *in fact* identical to the firing of C-fibers.[6]

Kripke, following Descartes, seems to rely on the idea that when you have a really "clear and distinct" idea you have access to how things are, metaphysically speaking. If one believes in this sort of access to metaphysical facts, it then makes sense to use the Kripke test, by which I mean the test that determines whether the imagined scenario can be explained away appropriately, to determine whether one has hold of a genuine metaphysical possibility or not. So, in the water/H_2O case, Kripke shows that, as it were, when your idea is made properly clear and distinct, you see that what you are really entertaining is the thought that something that behaves like water is not H_2O. Notice that the situation satisfying this description is indeed metaphysically possible. Since the same move doesn't work for the pain/C-fibers case, we conclude that there is a metaphysically possible world in which pain isn't the firing of C-fibers.

But suppose we reject the Cartesian model of epistemic access to metaphysical reality altogether. One's ideas can be as clear and distinct as you like, and nevertheless not correspond to what is in fact possible. The world is structured in a certain way, and there is no guarantee that our ideas will correspond appropriately. If one fol-

lows this line of thought, then the distinction Kripke points out between the pain/C-fibers case and the water/H_2O case turns out to be irrelevant to the question of what is or is not metaphysically possible. Thus, for all we know, pain *just is* the firing of C-fibers or, if functionalism is right, the realization of a certain functional state.

Early identity theorists, in their response to Cartesian conceivability arguments, protested that they only intended their theory to be empirical, and therefore it was not subject to objections from what was conceivable or not.[7] Kripke correctly pointed out the error of that sort of response. Empirical or not, if they were making identity claims, then a consequence of their theory is that it is not possible for some mental state not to be identical to its physical or functional correlate. But the basis of Kripke's objection lies in a strict distinction between metaphysical and epistemological possibility. Once we appreciate that distinction, the physicalist can return to her original ploy, i.e., to say that metaphysical consequences cannot be drawn from considerations of what is merely conceivable. Thus, without an argument to the effect that what is metaphysically possible is epistemologically accessible, the Cartesian argument fails.

Jackson's Argument

A similar problem—that is, a reliance on the Cartesian model of epistemic access to metaphysical reality or, in other words, using epistemoloical premises to support a metaphysical conclusion—seems to infect Frank Jackson's well-known "knowledge argument" against materialism. Jackson takes the thesis of physicalism to be the claim that "all (correct) information is physical information" (Jackson 1982, p. 127). of course, his use of the notion of information here is already fraught with ambiguity as between matters epistemological and metaphysical, a point to which I will return shortly. His argument against physicalism revolves around examples like the following:

> Mary is a brilliant scientist who is ... forced to investigate the world from a black and white room *via* a black and white television monitor. She specializes in the neurophysiology of vision and acquires ... all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like "red," "blue," and so on....
>
> What will happen when Mary is released from her black and white room or is given a color television monitor? Will she *learn* anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had *all* the physical information. *Ergo* there is more to have than that, and Physicalism is false. (Jackson 1982, p. 130)

There have been a number of replies to Jackson in the literature, and the sort of reply I am most interested in is exemplified by Horgan (1984). Horgan argues that Jackson is equivocating on the notion of "physical information." In one sense this might mean information expressed in terms used in the physical sciences. In another sense it might mean information about physical facts, processes, etc. It is only in the second sense that any reasonable physicalist is committed to the claim that all information is physical information. Of course, in this sense, the thesis could be better put by just saying that all token events and processes are physical events and processes—by which one means something like, they have a true description in the terms of the physical sciences. (Actually, I think any interesting doctrine of physicalism is committed to more than this, though it's difficult to pin down exactly how much more. At any rate, it doesn't affect the present point.) But no plausible version of physicalism is committed to the claim that all information is physical information in the first sense: in the sense that it is expressed in (or translatable into) the terms of the physical sciences.

What Mary's case shows, argues Horgan, is that there is information Mary acquires after leaving the room that isn't physical information in the first sense, but not that it isn't physical information in the second sense. Certainly, she

may think something like,"Oh, so *this* is what red looks like." Her experience of *learning* something new shows that she now knows something she didn't know before. She now knows what it's like to see red, which she didn't know before. But it doesn't follow that her new information isn't physical information in the second sense: that is, that it isn't information about a physical event or process. On the contrary, the case of Mary typifies the phenomenon of there being several distinguishable ways to gain epistemic access to the same fact. One cannot infer from a variety of modes of access to a variety of facts being accessed.

A similar emphasis on the distinction between the epistemology and the metaphysics of the matter underlies the following sort of reply to Jackson. What the case of Mary shows is that one can know which physical (or functional) description a mental state satisfies without knowing what it's like to occupy that state. But of course! After all, in order to know what it's like to occupy a state one has actually to occupy it! All Mary's newly acquired knowledge amounts to is her new experience, which is indeed new, since she didn't have those experiences until leaving the room. So it remains perfectly possible that what she learns is what it's like to occupy a certain physico-functional state. There is no threat to physicalism here.

Two Metaphysical Anti-Physicalist Replies

The common thread in the responses to both Kripke and Jackson is that their thought experiments demonstrate only an epistemological divide between different modes of access to what may, for all we know, be the very same phenomenon. On the one hand, we have certain physico-functional descriptions of certain states occupied by psychological subjects. On the other hand, we have whatever descriptions are derived from one's first-hand experience of these states. If these thought experiments show that physicalism leaves something out, it can't be in the sense that there are facts that physicalistic descriptions fail to pick out, since we have no argument to show that the two sorts of descriptions just cited do not refer to the same facts.

I will briefly consider two replies on behalf of the metaphysical anti-physicalist. First, perhaps Cartesian conceivability arguments can't demonstrate that qualia aren't physical states or processes, but they at least throw the burden of argument back onto the physicalist to show why we should think they are physical states or processes. The physicalist strategy presented above only opens a space for the physicalist hypothesis, but it doesn't give us any reason to believe it.

Fair enough. That's all it was intended to accomplish. The main burden of the physicalist argument is borne by considerations of causal interaction. If qualia aren't physical processes (or realized in physical processes), then it becomes very difficult to understand how they can play a causal role in both the production of behavior and the fixation of perceptual belief. Jackson himself admits the cogency of this argument, and therefore bites the bullet by endorsing epiphenomenalism. Those who don't find that bullet particularly appetizing must either show how the requisite mental-physical causal relations are possible on a dualist account or endorse physicalism.

The second reply on behalf of the metaphysical anti-physicalist goes like this.[8] Take some identity statement that is not epistemologically necessary, like (3) below:

3. The Morning Star = the Evening Star.

Though one might accept Kripke's claim that (3) is necessarily true if true at all, still one has to explain its apparent contingency. The way we do this is to say that what is contingent is that the very same heavenly body should appear where Venus does in the morning and also where it does in the evening. Notice that our explanation of the apparent contingency of (3) adverted to a real

distinction between two of Venus's properties: namely, appearing at a certain heavenly location in the morning and appearing at a certain heavenly location in the evening. That is, we can explain the epistemological state of conceiving of the Morning Star and the Evening Star as two distinct objects, despite their identity, by reference to two distinct properties through which we have epistemic access to the one object.

Suppose one grants that the absent qualia argument does indeed establish at least the epistemological possibility that a qualitative state and a functional state are distinct, even though they are in fact identical. In order to explain how it is possible to conceive of this one state as two distinct states, we must assume that there are (at least) two "modes of presentation" under which we apprehend this one state. Let us call them the "first-person mode of presentation" and the "third-person mode of presentation." But now we seem committed to the claim that there are at least two distinct properties of the state corresponding to the two modes of access, akin to the two spatiotemporal properties of Venus by which we gain epistemic access to it in the morning and in the evening. If so, this shows that qualitative character, the property by which we identify a conscious state in the first-person mode of access, is distinct from the property of playing a certain functional role, the property by which we identify that conscious state in the third-person mode of access. So, we seem to be back to deriving a metaphysical conclusion from an epistemological premise, namely, that the property of having a certain qualitative character is distinct from the property of playing a certain functional role (or being in a certain neurophysiological state).[9]

The physicalist, however, can reply as follows. Certainly whenever we conceive of a single object in two distinct ways—sufficiently distinct ways, in fact, that we believe we are conceiving of two distinct objects—the object in question must possess (at least) two distinct properties that correspond to these different modes of presentation. But whether or not we now have a problem for physicalism depends on which two district properties we find ourselves committed to. This requires some elaboration.

What the physicalist needs to maintain is that having a certain qualitative character is a physical or functional property. This reduction of qualitative character is necessary in order to account for the causal role that qualia play in the fixation of perceptual belief and the production of behavior. So, if the argument above could establish that having a certain qualitative character is a property distinct from a mental state's physical and functional properties, that would be the sort of metaphysical conclusion the anti-physicalist is after.

However, the argument above does not in fact establish the non-identity of having a certain qualitative character and any of a state's physical or functional properties. The argument begins with the premise that there must be two properties of the one state, providing two epistemic paths by which the subject conceives of that state, in order to account for the fact that it is epistemologically possible for someone to experience qualitative character without occupying the relevant physico-functional state. We can accept this premise and yet refuse to grant the conclusion—that having a certain qualitative character is irreducible to a state's physico-functional properties—by finding two other properties to provide the requisite epistemic paths. For instance, we can account for the conceivability of experiencing a certain quale without occupying the relevant physico-functional state by noting that the two relational properties, being thought of under the description "what I am now consciously experiencing" and being thought of under the description "the state that normally causes [such-and-such behavioral effects]," are not identical. However, there is no reason for the physicalist to claim that *these* two properties are identical, and therefore the argument above fails to mount a challenge to physicalism.

The Epistemological Reading

I have argued that on a metaphysical reading of "leave something out," Cartesian conceivability arguments cannot establish that physicalist theories of mind leave something out. However, there is also an epistemological sense of "leave something out," and it is in this sense that conceivability arguments, being epistemological in nature, can reveal a deep inadequacy in physicalist theories of mind.

For a physicalist theory to be successful, it is not only necessary that it provide a physical description for mental states and properties, but also that it provide an *explanation* of these states and properties. In particular, we want an explanation of why when we occupy certain physico-functional states we experience qualitative character of the sort we do. It's not enough for these purposes to explain the contribution of qualitative states to the production of behavior, or the fixation of perceptual belief; this is a job that a physicalist theory can presumably accomplish. (At least there is no reason stemming from conceivability arguments to suppose that it cannot.) Rather, what is at issue is the ability to explain qualitative character itself; why it is like what it is like to see red or feel pain.

Conceivability arguments serve to demonstrate the inability of physicalist theories to provide just this sort of explanation of qualitative character. To see this, consider again the disanalogy Kripke draws between statements (1) and (2) above. Kripke bases his argument on the fact that both statements appear contingent, and then distinguishes between them by pointing out that the apparent contingency of (1), but not of (2), can be explained away. My strategy is quite different. I see the disanalogy between the water/H_2O case and the pain/C-fibers case in the fact that there is an apparent *necessity* that flows from the reduction of water to H_2O, a kind of necessity that is missing from the reduction of pain to the firing of C-fibers.

The necessity I have in mind is best exemplified by considering statement (1′):

1′. The substance that manifests [such-and-such macro properties of water] is H_2O.

On Kripke's view, (1′) is in fact contingent, and it is the contingency of (1′) that explains the apparent contingency of (1). So, on his view, (1′) and (2) are on a par. Yet, it seems to me that there is an important difference between them. If we consider the apparent contingency that attaches to (2), we notice that it works in both directions: it is equally conceivable that there should exist a pain without the firing of C-fibers, and the firing of C-fibers without pain. However, the apparent contingency of (1′) only works in one direction. While it is conceivable that something other than H_2O should manifest the superficial macro properties of water, as Kripke suggests, it is not conceivable, I contend, that H_2O should fail to manifest these properties (assuming, of course, that we keep the rest of chemistry constant).

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-fibers: 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, that is, 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 vapor 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 vapor pressure. When the vapor 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?

One might object at this point that my position presumes something like the deductive-nomological account of explanation, an account that is certainly controversial.[10] In fact, I quite openly endorse the view that explanations involve showing how the explanandum follows from the explanans. I believe that the deductive-nomological model, in analyzing explanation in terms of exhibiting a necessary connection between explanans and explanandum, is certainly on the right track.

I am not committed, however, to the view that all explanations take the form of the "covering law" model described by Hempel (1965) in his classic account of explanation. For instance, Robert Cummins (1983) has argued that some explanations take the form of "property theories," in which the instantiation of one sort of property is explained by reference to the instantiation of some other properties. So we might, for example, explain a certain psychological capacity by reference to the physico-functional mechanisms that underlie it. In such cases we are not explaining one event by citing initial conditions and subsuming it under a law, so it does not quite fit the traditional deductive-nomological model.

I have no problem with Cummins's objection to the covering law model. Yet even in his example—explaining how a psychological capacity is instantiated by reference to the underlying mechanisms—the element of necessity is there, even if there is no subsumption under laws. For it is clear that if citing the relevant underlying mechanisms really does explain how the psychological capacity in question is instantiated, then it would be inconceivable that some creature should possess these mechanisms and yet lack the capacity. If not, if we could conceive of a situation in which a creature possessed the relevant underlying

mechanisms and yet didn't possess the capacity in question, then I would claim that we haven't adequately explained the presence of the capacity by reference to those mechanisms. For we are still left wondering what distinguishes the actual situation, in which the creature possesses the capacity, from those conceivable situations in which it (he/she) does not.

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 behavior. 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 being 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, and so forth—to H_2O.

But suppose someone pressed further, asking why being causally responsible for this particular syndrome of superficial properties should be so crucial.[11] 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 behavior 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 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.[12] To take another example, we say that one of water's superficial properties is that it is colorless. But being colorless is not a chemical property, so before we can explain why water is colorless in terms of the molecular structure of water and the way that such structures interact with light waves, we need to reduce colorlessness 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 color, 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 color.

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 analyzed 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.[13]

A Digression about Concepts

In order to clarify the sense in which it is inconceivable that something should be H_2O and not be water, I have had to slip into talking about concepts; even worse, talking about analyzing the contents of concepts. This is unfortunate for my position, since the whole topic of concepts is filled with controversy, and I do not yet see how to construct a theory of conceptual content that will do the work, briefly outlined above, that needs to be done. Let me briefly indicate where the problems lie.

In the literature on concepts and contents, various distinctions have emerged which are useful to our concerns here. First of all, we can distinguish between a concept's[14] "narrow content" and its "broad content."[15] A concept's broad content is its satisfaction conditions. This is the referential component of its content. The notion of narrow content is meant to capture that aspect of its content that is psychologically significant and independent of facts external to the subject. With regard to the famous Twin Earth example, narrow content is what my concept of water and my twin's concept of water have in common.

It is, of course, controversial whether or not it is narrow content that is relevant to the individuation of psychological states, or even whether there is such a thing as narrow content. However, I believe there is something psychologically significant that I and my twin have in common when we entertain the concept of water and, moreover, it is this aspect of our concept that seems relevant to the question of explanatory reduction. In will not defend this claim here, I will just presume it.

So, how do we characterize the narrow content of our concept of water? On one view, the "functional role" view,[16] narrow content is determined by the cluster of beliefs involving the concept of water that determine the inferential relations among them. On this view, to analyze the concept of water is just to present those central beliefs. Our concept of water is the concept of

a substance that..., where statements about the typical behavior of water fill in the blank. On a functional role view, then, to say that our concept of water can be analyzed as the concept of a causal niche is just to say that the beliefs which go into the blank all involve the causal role of water.

Thus a functional role view of narrow content seems quite amenable to my needs. However, there are real problems with this view. In particular, there is the problem of holism. That is, if you change any element of the description of causal relations definitive of the concept, you change the concept. Now, in the course of scientific investigation, we can expect to revise our beliefs about these causal connections as we learn more about the phenomenon under investigation. If such changes counted as changing the concept, then it wouldn't be *water* we were learning about when we discovered that water was H_2O. This would seem to be an intolerable consequence.

There are three ways one might deal with this problem. First, just bite the bullet and admit that our concepts are hopelessly holistic. Second, attempt to distinguish those elements of the functional role that are essential to the concept from those that are accidental, so that only changes in the essential elements constitute changes in concept. Third, find a different theory of narrow content.

One might argue that biting the bullet is not as bad as it seems on the grounds that we are only talking about *narrow* content. So long as we are atomistic about reference, we can still make sense of the claim that two theories contain conflicting claims *about water*, since they are talking *about* the very same thing. However, what we are looking for here is a notion of conceptual content suitable for grounding the explanatory relation, and this is clearly a matter of narrow content. Unless we can build stability into the notions, it is unclear how to make sense of the idea that the chemical theory of water explains why water behaves the way it does.

The second sort of strategy has a sad history, and I do not see how to make it work. For any element of the functional role you pick as essential, there always seems to be a story you can tell in which that element is missing and yet it seems intuitively right to claim that the subject still has the concept in question. As for a different theory of narrow content, the only one I know, that departs radically from the functional role theory, is Fodor's (1987) theory of narrow content as a function from contexts to broad contents. It is not at all clear to me how the notion of a conceptual content as the specification of a causal niche could be made to work on this view.

To sum up, there seems to be a need for a theory of conceptual content that both grounds explanatory reductions on the basis of some sort of functional/causal analysis of the requisite concepts, and yet does not entail holism. I do not have such a theory, and so must content myself with merely characterizing this desideratum on a theory yet to be developed.

Qualia Again

If we apply the same model of explanatory value to the theoretical reduction of qualia as we used for the reduction of water, then we need to look for a property that is being reduced and then a property, or set of properties, by which the to-be-reduced property is normally picked out. Of course, this raises a problem. When it comes to something like the qualitative character of a sensation of red, what other property could we point to to play the role of the reference-fixer? We seem to pick out this property by itself. The distinction between the property to be reduced and the properties by which we normally pick it out, or its superficial manifestation, seems to collapse. (Obviously this is connected to Kripke's point about the appearance/reality distinction not getting a hold in this case.)

There are, of course, other properties of qualia that we can expect a theoretical reduction to explain; namely, those properties associated with their causal role in mediating environmen-

tal stimuli and behavior. It is precisely on the grounds that a particular physico-functional property can explain the "behavior" of qualitative states that we would endorse an identification between a particular quale and that property. Furthermore, if that were all there were to our concept of qualitative character—as the analytical functionalist maintains—then there would be no difference between the theoretical reduction of water and of qualia with respect to explanatory success. But the very fact that one can conceive of a state playing that role and yet not constituting a qualitative experience shows, or at least so I have argued, that causal role is not all there is to our concept of qualitative character.

What seems to be responsible for the explanatory gap, then, is the fact that our concepts of qualitative character do not represent, at least in terms of their psychological contents, causal roles. Reduction is explanatory when by reducing an object or property we reveal the mechanisms by which the causal role constitutive of that object or property is realized. Moreover, this seems to be the only way that a reduction could be explanatory. Thus, to the extent that there is an element in our concept of qualitative character that is not captured by features of its causal role, to that extent it will escape the explanatory net of a physicalistic reduction.

Conclusion

I will conclude by drawing out another consequence of this discussion of the explanatory gap. It is customary to attack the mind-body problem by a divide-and-conquer strategy. On the one hand, there is the problem of intentionality. How can mere matter support meaning; how can a bit of matter be *about* something? On the other hand, there is the problem of consciousness, or, to be more specific, the problem of qualitative character. How can there be something it is like to be a mere physical system? By separating the two questions, it is hoped that significant progress can be made on both.

Certainly in recent years we have come to have a deeper understanding of the issues surrounding the question of intentionality, and this progress has been largely the result of divorcing the question of intentionality from the question of consciousness.[17] However, if I am right, there may be more of a connection between the problem of qualitative character and the problem of intentionality than it is fashionable now to suppose. It is not that one needs to be capable of experiencing qualia in order to bear intentional states, as Searle would have it. Rather, since the problem of qualitative character turns out to be primarily epistemological, the source of which is to be found in the peculiar nature of our cognitive representations of qualitative character, a theory of intentional content ought to explain what makes these representations so uniquely resistant to incorporation into the explanatory net of physical science.[18] Thus the problem of qualia threatens to enlarge into the problem of the mind generally.

Acknowledgments

An earlier version of this chapter was delivered at the conference on Mind, Meaning, and Nature, at Wesleyan University, March 31, 1989. The chapter was completed while I was holding a fellowship from the National Endowment for the Humanitites. I would also like to thank Louise Antony, David Auerbach, Martin Davies, and Georges Rey for helpful discussions and critical comments on earlier drafts.

Notes

1. See Levine (1983) where I first argued for the existence of an explanatory gap.

2. Of course, it is a non-trivial question to decide which predicates count as "physical" predicates, but for present purposes we need not attempt a precise explication of the notion.

3. As Kripke (1980, p. 154) puts it, there is a "certain obvious element of contingency" about such theoretical identity statements.

4. For extensive discussion of the absent and inverted qualia hypotheses, see Block and Fodor (1972); Block (1978, 1980); Horgan (1984); Shoemaker (1984, chs. 9, 14, and 15); Conee (1985); Levine (1989).

5. For various versions of this position, see Block (1978, 1980), Horgan (1984), and Shoemarker (1984).

6. Given that Kripke is largely responsible for drawing the philosophical world's attention to the distinction between epistemological possibility and metaphysical possibility, it might seem odd to accuse him of confusing the two in this case. I diagnose his mistake as follows. Since he believes that any state which appears painful is thereby a pain, he infers that there is no appearance/reality distinction with respect to pain, and therefore epistemological and metaphysical possibility collapse in this case. But even if he's right that any state that appears to be a pain is a pain, he still has to justify the premise that it's possible for one to suffer even apparent pain without having one's C-fibers firing, and he can't do that, I contend, merely by noting that it *seems* possible.

7. See, for instance, Smart's (1959) reply to his "Objection 2."

8. This objection was suggested to me by a discussion in White (1986). The analogy to the Morning Star–Evening Star case is his.

9. As White expicitly acknowledges, a precursor of this objection can be found in Smart's (1959) famous "Objection 3."

10. For the classic presentation of the deductive-nomological model of explanation, see Hempel (1965, ch. 12).

11. 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*.

12. 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.

13. To a certain extent my argument here is similar to Alan Sidelle's (1989) defense of conventionalism, though I don't believe our positions coincide completely.

14. Some readers might find my speaking of a *concept's* content, as opposed to a *term's* content confusing. I am interested in the nature of our thoughts, not with their expression in natural language. For present purposes, we can think of a concept as a term in whatever internal, mental language is employed in our cognitive processing.

15. For the source of this distinction, see Putnam (1975). For further discussion of its significance for psychology, see Fodor (1987, ch. 3) and Burge (1986).

16. See Block (1986) for a defense of the functional role view.

17. For a dissenting opinion on the question of divorcing intentionality from consciousness, see Searle (1989).

18. See Rey (chapter 12) and Van Gulick (chapter 7) in Davies and Humphries for suggestive approaches to just this problem.

References

Block, N. (1978). Troubles with functionalism. In C. Wade Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology*. Vol. 9, *Minnesota Studies in the Philosophy of Science*. Minneapolis: University of Minnesota Press.

———. (1980). Are absent qualia impossible? *Philosophical Review* 89, 257–74.

———. (1986). Advertisement for a semantics for psychology. In P. A. French, T. E. Uehling, Jr., and H. K. Wettstein (eds.), *Studies in the Philosophy of Mind*. Vol. 10, *Midwest Studies in Philosophy*. Minneapolis: University of Minnesota Press, 615–78.

Block, N., and Fodor, J. A. (1972). What psychological states are not. *Philosophical Review*, 81, 159–81.

Burge, T. (1986). Individualism and psychology. *Philosophical Review* 95, 3–46.

Conee, E. (1985). The possibility of absent qualia, *Philosophical Review* 94, 345–66.

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Davies, M., and Humphreys, G. (eds.). (1993). *Consciousness*. Oxford: Blackwell.

Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press

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

Horgan, T. (1984). Functionalism, qualia, and the inverted spectrum. *Philosophy and Phenomenological Research* 44, 453–69.

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly* 32, 127–36.

Kripke, S. (1980). *Naming and Necessity*. Oxford: Blackwell.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64, 354–61.

———. (1989). Absent and inverted qualia revisited. *Mind and Language* 3, 271–87.

Putnam, H. (1975). The meaning of meaning. In K. Gunderson (ed.), *Language, Mind, and Knowledge*, 131–93. Minneapolis: University of Minnesota Press.

Rey, Georges (1993). Sensational sentences. In Davies, M., and Humphreys, G. (eds.) *Consciousness*, 240–57. Oxford: Blackwell.

Searle, J. (1989). Consciousness, unconsciousness, and intentionality. *Philosophical Topics* 17, 193–209.

Shoemaker, S. (1984). *Identity, Cause, and Mind*. Cambridge: Cambridge University Press.

Sidelle, A. (1989). *Necessity, Essence, and Individuation: A Defense of Conventionalism*. Ithaca: Cornell University Press.

Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review* 68, 141–56.

Van Gulick, Robert (1993). Understanding the phenomenal mind: Are we all just armadillos? In Davies, M., and Humphreys, G. (eds.), *Consciousness*, 134–54. Oxford: Blackwell.

White, S. L. (1986). Curse of the qualia. *Synthese* 68.