Theories of Consciousness

David Papineau

Introduction

My target in this paper is "theories of consciousness". There are many theories of consciousness around, and my view is that they are all misconceived. Consciousness is not a normal scientific subject, and needs handling with special care. It is foolhardy to jump straight in and start building a theory, as if consciousness were just like electricity or chemical valency. We will do much better to reflect explicitly on our methodology first. When we do this, we will see that theories of consciousness are trying to answer a question that isn't there.

Consciousness as a Determinable Property

Let me begin with a useful distinction. We can think of consciousness as a determinable property, whose determinates are more specific modes of consciousness like being in pain, tasting chocolate, seeing an elephant and so on. By way of analogy, contrast the determinable property, having a shape, with the more specific (determinate) properties square, triangular, elliptical. Or again, contrast the determinable property being a car, with the determinates being a Ford, a Rover, a Rolls-Royce, and so on. The idea here is simply the notion of a genus, which then divides into a number of more restrictive species. In this way, then, being conscious is a general property, whose instances then all have some more determinate conscious feature like being in pain. (1)

The theories of consciousness which are my target in this paper are theories of the determinable property of consciousness-as-such, not determinate conscious properties like pain, or seeing an elephant, or whatever. Many theories of just this kind are on offer nowadays. Thus, to pick a quick sample, consider the identification of consciousness with quantum collapses in cellular microtubules (Penrose, 1994), or with operations in the global workspace (Baars, 1988) or with competition for action control (Challice, 1988), or with informational content (Chalmers, 1996, Tye, 1995, Dretske, 1995) or, again with, higher-order thought (Armstrong, 1968, Rosenthal, 1996, Lycan, 1996, Carruthers, forthcoming). These are all theories of what it takes to be conscious at all, not of more determinate states like feeling a pain, and so on. My argument will be that theories of this kind are barking up the wrong tree.

Before I get on to theories of the determinable property, being conscious, it will be useful first to explain at some length how I think of determinate conscious properties, as my analysis of general "theories of consciousness" will be informed by this understanding.

Determinate Conscious Properties are Physical Properties

I am a physicalist about determinate conscious properties. I think that the property of being in pain is identical to some physical property. This is not because I want to be provocative, or because I am caught up by current philosophical fashion, but because I think that there is an overwhelming argument for this identification, namely, that conscious mental states would have no influence on our behaviour, which they clearly do, were they not identical with physical states.

This is not the place to analyse this argument in any detail. But three quick comments will be in order.

First, you might want to ask, if so simple an argument can establish physicalism, why everybody hasn't always been persuaded by it? My answer is that it is a simple argument, but that until recently a crucial premise was not available. This is the premise that physical effects, like behaviour, are always fully determined, insofar as they are determined at all, by prior physical causes. This is a highly empirical premise, and moreover one which informed scientific opinion didn't take to be fully evidenced until some time into this century. It is this evidential shift, and not any tide of philosophical fashion, which has been responsible for the recent rise of physicalism. (For more details on this history, see Papineau, 2000.)

Second, let me say something about what I mean by "physical". I am happy to leave this quite vague in this paper. In particular, I would like it to be read in such a way as to include "functional" properties (like having-some--physical-property-produced-by-bodily-damage-and-causing-avodiance-behaviour), and physiological properties (like having your C-fibres firing) as well as strictly physical properties like mass, position and quantum collapses. This is skating over a number of tricky issues (in particular the issue of whether functional and other higher-level properties can themselves really cause behaviour, if that behaviour always has full physical causes). Fortunately this issue is orthogonal to my concerns here. The important point for present purposes is only that I want to identify determinate conscious properties with indepedently identifiable properties of a general scientific-causal sort, with properites that aren't sui generis irreducibly conscious properties. Given this, it doesn't matter here exactly which kind of properties we count as broadly "physical" in this sense. (For more on this, see Papineau, 1998.)

Third, and to forestall any possible confusion, I would like to emphasise that physicalism does not deny the "what-it's-likeness" of conscious occurrences. To say that pain is identical with a certain physical property is not to deny that it is like something to be in pain. Rather, it is to affirm that it is like something to be in a certain physical state. Of course it is like something to experience pain, or to see red, or to taste chocolate. And these experiences matter, especially to their subjects. But, insists the physicalist, they are not non-physical things. They are just a matter of your having some physical property. They are how it is for you, if you have that physical property.

Phenomenal Concepts and Third-Person Concepts

While I am a physicalist about determinate conscious properties, I am a sort of dualist about the concepts we use refer to these properties. I think that we have two quite different ways of thinking about determinate conscious properties. Moreover, I think that is crucially important for physicalists to realize that, even if conscious properties are just physical properties, they can be referred to in these two different ways. Physicalists who do not acknowledge this, and there are some, will find themselves unable to answer some standard anti-physicalist challenges.

I shall call these two kinds of concepts "phenomenal" concepts and "third-person" concepts. The idea, then, is that we have two quite different ways of thinking about pain, say, or tasting chocolate, or seeing an elephant, both of which refer to the same properties in reality. By way of obvious analogy, consider the case where we have

two names, "Judy Garland" and "Frances Gumm", say, both of which refer to the same real person.

The distinction between concpets is similar to David Chalmers' distinction between "phenomenal" and "psychological" concepts. The reason I have contrasted phenomenal concepts with "third-person concepts" rather than with Chalmers' "psychological concepts" is that I am not currently concerned to analyse our non-phenomenal ways of thinking about conscious states in any detail (though I will say a bit more on this below). Chalmers has specific views on this matter, which make him focus in the first instance on functional concepts mental states, which he calls "psychological" concepts (for example, the concept of having-some-physical-property-produced-by-bodily-damage-and-causing-avodiance-behaviour). While such psychological concepts are one instance of what I mean by "third-personal" concepts, I want also to include here any other concepts which identify their referents as parts of the third-personal, causal world, including physiological concepts (C-fibres firing) strictly physical concepts (involving ideas of mass, position, and so on), and indeed everyday descriptive concepts like "his reaction to that bad news last Thursday" or "the causes of his offensive behaviour".

Third-personal concepts will thus be a broad category, but that doesn't matter, given that we don't really need a defintion of the category beyond its contrast with "phenomenal" concepts. These comprise the more interesting category. When we use phenomenal concepts, we think of mental properties, not as items in the third-personal causal world, but in terms of what they are like. Consider what happens when the dentist's drill slips and hits the nerve in your tooth. We can think of this third-personally, in terms of nerve messages, brain activity, involuntary flinching, and so on. Or we can think of it in terms of what it would be like, of how it would feel if that happened to you.

How Physicalists Can Stop Worrying and Learn to Love Phenomenal Concepts

Phenomenal concepts are normally introduced by anti-physicalist philosophers, by philosophers who want to resist the identification of phenomenal properties with physical properties. Such anti-physicalists aim to move, from the existence of distinctive non-physical ways of thinking, to the existence of distinctive non-physical ways of being.

Now, some physicalists aim to resist this move by denying the existence of phenomenal concepts, by denying that there are any distinctive non-physical ways of thinking (Dennett, 1991, Churchland and Churchland, 1998). But this seems to me quite the wrong move. There is nothing in phenomenal concepts per se to worry the physicalist. In particular, they don't entail that there are distictive phenomenal properties.

A good way to show this is to consider Frank Jackson's story of Mary (Jackson, 1986). Jackson takes this story to demonstrate the existence of distinctive phenomenal properties. But the physicalist can respond that, while it is certainly a good way of demonstrating that there are distinctive phenomenal concepts, the further move to non-physical properties is invalid.

In Jackson's thought-experiment, Mary is an expert on colour vision. She knows everything there is to know about the physics and physiology of peoples' brains when

they see colours. However, Mary is peculiar in that she has never seen any colours herself. She has always lived in a house with an entirely black-and-white interior, she doesn't have a colour television, all her books have only black-and-white illustrations, and so on. Then one day Mary goes out and sees a red rose.

At this point, argues Jackson, she learns about something she didn't know before. As we say, she now "knows what it is like to see red". And since she already knew about everything physical connected with colour experience, continues Jackson, this must involve her now knowing about some distinctive phenomenal property associated with red experiences, which she didn't have access to before she saw the rose.

However, as I suggested, physicalists who recognize phenomenal concepts needn't accept this argument. For they can respond that, while there is indeed a genuine before-after difference in Mary, this is just a matter of her acquiring a new concept of seeing red. The property she refers to with this concept is still a perfectly good physical property, the physical property, whatever it is, that is present in just those people who are seeing red, and which she could think about perfectly well, albeit using third-personal concepts, even before she saw the rose. (In the terminology of philosophical logic, we can say that Mary has a new Fregean thought, but not a new Russellian one.)

To fill out this suggestion, note that the essential change in Mary, now that she "knows what it is like to see red", involves two things. First, she is now able to have memories, and other thoughts, which imaginatively recreate the experience, as when she recalls what it was like to see red, or anticipates what it will be like to see red. Second, she is also now able introspectively to reidentify other experiences as of that type, as when she thinks that she is now having the same experience as when she saw the rose.(2)

When I speak of Mary's acquiring a new phenomenal concept, I mean she is able to think new (Fregean) thoughts by using these new powers of recreation and reidentification. Thus she will be able imaginatively to recreate the experience in question, and thereby think such thoughts as "Having a Ø [and here she imagines the experience] won't be possible for people who are colour-blind", or "Jim will have a Ø in a minute". Again, when she is actually having the experience in question, she will be able to think thoughts like "This Ø [here she identifies an aspect of her current experience] is an hallucination caused by what I ate", or "Looking at that square has made me have this Ø afterimage". Thoughts of these kinds seem quite unproblematically truth-evaluable, and in particular there seems no special difficulty in understanding the contribution that the Ø-element makes to the truth conditions of these thoughts. So we can think of the Ø-element as a concept, in the familiar sense of an item that makes a sytematic contribution to the truth conditions of the thoughts it enters into.

Now, as I said, Jackson and others take the existence of Mary's new phenomenal concept, and of phenomenal concepts in general, to imply the existence of distinctive phenomenal properties. The idea, presumably (though this is not often spelled out), is that the before-after difference in Mary (she now "knows what it's like to see red", when before she didn't) somehow derives from her new-found acquaintance with the phenomenal features of her experience. On this model, then, the possession of a phenomenal concept requires that the possessor previously be directly acquainted with

some phenomenal property. This is why nobody can "know what an experience is like" prior to having it.

However, given the suggestions I have made about the structure of phenomenal concepts, there is an obvious alternative physicalist story to be told. This which accounts equally well for the fact that you can't "know what an experience is like" prior to having it, and does so without invoking any special phenomenal properties.(3)

Here is the obvious physicalist explanation. Suppose that imaginative recreation depends on the ability to reactivate (some of) the same parts of the brain as are activated by the original experience itself. Then it would scarcely be surprising that we can only do this with respect to types of experience we have previously had. We can't form replicas, so to speak, if external stimulation hasn't fixed a mould in our brains. Less metaphorically, we can only reactivate just the parts of the brain required for the imaginative recreation of some experience E if some actual instance of E has previously activated those parts. Similarly, suppose that introspective identification of some experience requires that it is compared with some model or template stored in the brain. Again, it would scarcely be surprising that we should need an original version of the experience, in order to form the template for such comparisons.(4)

So this now gives us a physicalist account of what is involved in Mary's coming to "know what it is like" to see red. This account acknowledges that Mary acquires a new phenomenal concept of seeing red. But it denies that this new concept points to any new non-physical property. The change in Mary does not involve any acquaintance with a phenomenal property. Rather, her brain is lastingly altered in certain ways, and this now allows her imaginatively to recreate and introspectively to reidentify an experience she could previously only think about in a third-person way. Seen in this way, it is clear there is nothing in the idea of phenomenal concepts as such which bars them to physicalists.

Kripkean Intuitions

Not only can physicalists happily accept the existence of distinctive phenomenal concepts, but they should accept them, otherwise they will have trouble responding to Saul Kripke's famous argument against mind-brain identity.

At its simplest, Kripke's argument starts from the imaginability of zombies. Surely it makes sense to suppose that there could be a being physically just like me, but with no feelings, an unconscious automaton. But if zombies are imaginable, then they are possible. And, if they are possible, then it would seem to follow that conscious properties are distinct from physical properties, for it is precisely their lack of conscious properties, despite their sharing all our physical properties, that creates the possibility of zombies.

Physicalists should object to the slide from imaginability to possibility. Zombies may be imaginable, but they aren't possible. Even God could not make a zombie. Since my conscious properties are nothing but a subclass of my physical properties, any being that shares all my physical properties will therewith share all my conscious properties.

But if zombies aren't possible, as the physicalist must say, then how come they are imaginable, as clearly they are? This is where Kripke's argument bites. A natural explanation for our apparent ability to imagine many impossibilities (such as the

impossibility that H2O is not water) is that we are picking out an entity (water) via some contigent features it has in this world (odourless, colourless, etc), and then imagining the genuinely possible world in which H2O (that is, water) does not have those contingent features. But no strategy like this is going to help physicalists with the mind-brain case, For if they try arguing that, in imagining zombies, we are imagining beings who really do have pain, and only lack those properties by which we pick out pain in this world (hurtfulness, perhaps, or achiness), then they can be challenged to explain how the physically identical zombies can lack these further properties, unless these properties are themselves non-physical.

We can see why physicalists get into trouble here. In effect, the imaginability of zombies shows that we have a concept of pain which is different from any possible third-personal concept: we can conceive of a being who does not satisfy the pain concept, however many third-personal concepts it satisfies. The water-H2O analogy then invites the physicalist to explain the distinctive nature of this pain concept in terms of certain distinctive properties (hurtfulness, achiness) by which it picks out its referent. However, if this invitation is accepted, then the physicalist runs into trouble. For it seems that these supposed distinctive properties will only do the job of explaining why the pain concept is distinct from all third-personal concepts if they are themselves non-physical properties. After all, if the pain concept referred by invoking physical properties, this would seem to imply that it must be the same as some ordinary third-personal concept, which it isn't.

There may seem to be a loophole here. Couldn't physicalists still argue that the properties invoked by the pain concept are physical, but that they are invoked using distinctive phenomenal concepts ("hurtfulness", achiness"). But this thought leads nowhere. For now their opponents can mount just the same challenge to the concepts of hurtfulness or achiness as they originally applied to the concept of pain. If "hurtfulness" is not true of our possible physical duplicates, then it must be distinct from any third-personal concept, and so phsyicalists once more owe some explanation of how it can be so distinct, and we are back just where we started.

Physicalists shouldn't accept the invitation implict in the water-H2O analogy in the first place. That is, they shouldn't aim to explain the distinctive nature of the phenomenal pain concept in terms of its reference being fixed by invocation of distinctive properties. Instead they should argue that such concepts refers to their objects directly, without invoking any distinctive properties, and that their distinctive nature lies elsewhere, in the fact that their deployment involves exercises of imagination or introspection.

If this strikes you as ad hoc, note that some concepts of properties must refer to their objects directly, without invoking other properties, on pain of regress. It can't be that every concept of a property refers to that property as "the property which has some other property F", or in any similar other-F-invoking way, since the concept which refers to this other property F will then need to invoke some further property G, and so on. So there must be some concepts of properties that refer directly, at least in the sense that they don't do so by invoking other properties. I do not claim that phenomenal concepts are the only such directly-referring concepts of properties. In fact I think it likely that there are many such concepts. But all that matters for the moment is that there must be some, and that phenomenal concepts will feature among them.(5)

Some unfinished business remains. If phenomenal concepts pick out their referents without invoking contingent properties of those referents, then, once more, how do physicalists explain the imaginability of zombies? What exactly is the content of our thought when we imagine a physical duplicate which does not feel pain? The water-H2O model is no longer available. Physicalists can't now say that in imagining all our physical properties, and yet no pains, we are imagining a world in which those physical properties lack the contigent features by which we pick out pains. For we have now agreed that pain isn't picked out by any contingent features.

However, there is an obvious enough alternative solution. Instead of trying to identify some genuine possibility which we are imagining, physicalists can simply say that there is no real possibility associated with the thought that pains are not C-fibres firing (or any other physical property), and that the thinkability of this thought consists in nothing beyond the facts that we have a concept pain, a concept C-fibres firing, the concepts are and not, and the power to form a thought by joing them together.

Indeed, having come this far, we can see that we may as well have said the same thing about imagining the impossibility that H2O is not water. There is no real need to tell the complicated Kripkean story about our really imagining something else, namely, a world in which H2O (that is, water) lacks the properties which fix reference to water in this world. Why not simply say that "water" and "H2O" are different concepts, which they clearly are for most people, and use this fact alone to explain how those people can, without conceptual inconsistency, think the "impossible" thought that H2O is not water. The point is that there is nothing difficult about thinking an impossible thought, once you have two terms for one thing. Just join them in a thought where they flank a term for non-identity, and there you are.

Of course, there remains a genuine disanalogy between the H2O-water case and the mind-brain cases. Since "water" still does arguably refer by invoking properties, there indeed is a genuine possibility in the offing here (the possibility that H2O not be odourless, etc), even if we don't need this possibility to provide a content for "H2O =/ water" thoughts. By contrast, there is no genuine possibility corresponding to the thought that zombies might have no feelings. Since phenomenal concepts don't refer by invoking distinctive conscious properties, there is simply no possibility at all corresponding to the thought that a being may share your physical properties yet lack your conscious ones.

The Antipathetic Fallacy

There is a further reason why physicalists will do well to recognize a distinctive species of phenomenal concepts. It will help to explain why physicalism seems so implausible.

For there is no denying that intuition weighs against physicalism. I pointed out earlier that there is a strong argument for identifying conscious properties with physical properties, namely, that modern science shows that this is the only way of respecting the casual significance that we ordinarily ascribe to conscious states. Still, it is striking that, even in the face of this argument, many people continue to find it unbelievable that conscious states should be identical with physical states. This reaction contrasts with the response to other theoretical identifications. Thus, it is in a way surprising that water turned out to be H2O, or heat to be molecular motion. But few people continue to resist these conclusions, once they appreciate the evidence.

Mind-brain identification is different, in that intuition continues to object, even after the evidence is on the table. How can pain (which hurts so) possibly be the same thing as insensate molecules rushing around in nerve fibres? Or again, as Colin McGinn is so fond of asking, how can our vivid technicolour phenomenology (our experience of reds and purples and so on) possibly be the same as cellular activity in squishy gray matter? (McGinn, 1991.)

The difference between phenomenal concepts and third-personal concepts yields a very natural explanation of these anti-physicalist intuitions. Consider the two ways in which phenomenal concepts can be deployed, that is, in imaginative recreations and in introspective identifications. Both these exercises of phenomenal concepts have the unusual feature that we effectively use the experiences being referred to in the act of referring to them. When we imaginatively recreate an experience, we activate a faint copy of the original experience (cf Hume on ideas and impressions), and when we reidentify an experience, we think by bringing an actual experience under some comparison.

In both these cases the experience itself is in a sense being used in our thinking, and so is present in us. For this reason exercising a phenomenal concept will feel like having the experience itself. When you imagine a pain, or seeing red, or even more when you attend to these experiences while having them, versions of these experiences themselves will be present in you, and because of this the activity of thinking about pain or seeing red will introspectively strike you as involving the feeling of these experiences themselves.

Now compare exercises of some third-personal concept which, according to the physicalist, refers to just the same state. No similar feelings there. To think of C-fibres firing, or of some-physical-state-which-causes-damage avoidance, doesn't in itself create any feeling like pain. Or, again, thinking of grey matter doesn't in itself make you experience colours.

So there is a intuitive sense in which exercises of third-personal concepts "leave out" the experience at issue. They "leave out" the pain and the technicolour phenomenology, in the sense that they don't activate or involve these experiences. Now, it is all too easy to slide from this to the conclusion that, in exercising third-personal concepts, we are not thinking about the experiences themselves. After all, doesn't this third-personal mode of thought "leave out" the experiences, in a way that our phenomenal concepts do not? And doesn't this show that the third-personal concepts simply don't refer to the experiences denoted by our phenomenal concept of pain?

This line of thought is terribly natural, and I think it is largely responsible for widespread conviction that the mind must be extra to the brain. (Consider again the standard rhetorical ploy: "How could this panoply of feeling arise from mere neuronal activity?") However, this line of thought is a fallacy (which elsewhere I have dubbed the "antipathetic fallacy"). There is a sense in which third-personal concepts do "leave out" the feelings. Uses of them do not in any way activate the experiences in question, by contrast with uses of phenomenal concepts. But it simply does not follow that third-personal concepts "leave out" the feelings in the sense of failing to refer to them. They can still refer to the feelings, even though they don't activate them.

After all, most concepts don't use or involve the things they refer to. When I think of being rich, say, or having measles, this doesn't in any sense make me rich or give me measles. In using the states they refer to, pheneomenal concepts are very much the exception. So we shouldn't conclude on this account that third-personal concepts, which work in the normal way of most concepts, in not using the states they refer to, fail to refer to those states.

This then offers a natural account of the intuitive resistance to physicalism about conscious experiences. This resistance arises because we have a special way of thinking about our conscious experiences, namely, by using phenomenal concepts. We can think about our conscious experience using concepts to which they which bear a phenomenal resemblance. And this then creates the fallacious impression that other, third-personal ways of thinking about those experiences fail to refer to the felt experiences themselves.(6)

Implicit Dualism

Let me now return to the main topic of this paper, "theories of consciousness", in the sense of theories of consciousness-as-such, of the determinable property of consciousness, rather than its determinates.

At first pass, I would say that much theorising of this kind is motivated by more or less explicit dualism. Go back to determinate mental states like pain, seeing red, and so on, for a moment. If you are a dualist about such states, that is, if you think that in addition to their physical underpinnings these states also involve some distinct non-physical property, floating above the physical, as it were, then you will of course think that there is something terribly important common to all conscious states. They involve a special kind of non-physical property not found in the rest of the natural world. And if you think this, then of course you will want a theory about these special non-physical goings-on, a theory that tells you about the kinds of circumstances will generate these extra non-physical states.

Some of those who trade in theories of consciousness are quite overt about their dualist motivations. David Chalmers, for example, argues explicitly that conscious properties are extra to any physical properties, and so actively urges that the task of a "theory of consciousness" is to figure out which physical process give rise to this extra realm. He compares the theory of consciousness with the nineteenth-century theory of electromagentism. At one time it had been supposed that electromagnetism could be explained in terms of more basic mechanical processes. But James Clerk Maxwell and his contemporaries realized that this was impossible, and so added electromagnetism to the list of basic elements of reality. Chalmers urges exactly the same move with respect to consciousness. We need to recognize conscious experience as an additional feature of nature, and figure out the theoretical principles governing its generation.

Not all theorists of consciousness are as upfront as Chalmers. Yet the same commitments can be discerned even among thinkers who would be disinclined to consider themselves dualists. Thus theorists who begin by explicitly disavowing any inclinations towards dualism will often betray themselves soon afterwards, when they start talking about the physical processes which "generate" consciousness, or "cause" it, or "give rise to" it, or "are correlated with" it. These phrases may seem innocuous, but they implicitly presuppose that conscious properties are some extra feature of

reality, over and above all its physical features. That they come so readily to thinkers who do not think of themselves as dualists only testifies to the strength of anti-physicalist intuition. You may recognize the theoretical difficulties which accompany dualism, and wish sincerely to avoid them. But the peculiar structure of phenomenal concepts will grip you once more, and persuade you that third-personal ways of thinking inevitably "leave out" the crucial thing. So conscious feelings can't just be physical states, but must in some sense "arise from" them, or be "generated by" them. And then of course it will seem obvious, as before, that we need a "theory of consciousness". For what could be important than to figure out which physical processes have the special power to "generate" consciousness?

In this paper I shall have no further interest in theories of consciousness motivated in this way. I take the points already made in earlier sections to show what is wrong with dualism, and therewith to discredit the enterprise of finding out which physical states "give rise" to some extra realm of conscious being. There is no such extra realm, and so any theory seeking to identify its sources is embarking on a wild goose chase.

Physicalist Theories of Consciousness

Rather, what I shall consider from now on is whether there is room for theories of consciousness within a serious physicalism which identifies determinate conscious properties with physical properties, and does not slip back into thinking of the physical properties as "giving rise" to the conscious ones. Certainly there are plenty of serious physicalists who defend this possibility. They are quite clear that conscious properties are one and the same as physical properties, yet still want a theory that will tell us what is common to all cases of consciousness.

But I have severe doubts. I think that once we give up on dualism, the motivation for theorising of this kind disappears. When we follow the argument right through, and make sure that dualists thoughts are not allowed to intrude anywhere, then it will become unclear what such theories of consciousness-in-general are trying to do.

This conclusion is by no means obvious. The idea of a physicalist theory of consciousness-as-such certainly makes initial sense. It is perfectly normal for a scientific theory to identify the physical property which constitutes the real nature of some everyday kind. Thus science has shown us that water is H2O, and that genes are sequences of DNA, and many other such things. So why shouldn't it show us which physical property constitutes the real nature of consciousness?

Physicalists can find another good model in nineteenth-century physics, to set against Chalmers' appeal to Maxwell's theory. Where Chalmers appeals to electromagnetism, they can appeal to temperature. In the case of temperature, physics went the other way. Instead of adding temperature to the fundamental components of reality, it explained it in terms of a more basic mechanical quantity, namely mean kinetic energy. Similarly, argue physicalists about consciousness-as-such, we need a scientific theory that will identify the underlying physical property common to all cases of consciousness, and thereby show us what consciousness really is.

However, I don't think that this programme can be carried through. This is because I am doubtful about the concept of consciousness-as-such, the concept of a state's being like something. I don't think that this notion succeeds in picking out any kind. So there is no possibility of a reductive scientific theory which identifies the essence of

this kind. Such a theory will lack a target. We think our concept of consciousness gives us a good grasp of a real kind, but in fact there is nothing there.

At first this idea may seeem absurd. What could be more obvious than the difference between states it is like something to have, and those which are not? Am I a zombie, that I don't know? But I would ask readers to bear with me. The notion of a state's "being like something" is not a normal concept, and we shouldn't take it for granted that it works like other concepts.

Perhaps I should make it clear that I do not want to deny that there are certainly plenty of mental states which are like something for human beings, and plenty of other states which are not. But we need to treat the language involved in this claims with caution. I shall argue that the form of words, "being like something", does not draw a line across the whole of reality, with the states that are like something on one side, and those that aren't on the other.

I am not going to try to convince you of this head-on, however. My strategy will be to creep up on this conclusion from behind, by considering the methodology adopted by those who trade in theories of consciousness. At first sight, these theorists look as if they are simply trying to do for consciousness what science has done for water or temperature. But when we look more closely at the precise methodology adopted by these theorists, we will find that it doesn't really add up. This will lead me to reflect on the notion that defines the object of such theorising, the notion of consciousness-as-such, and to my eventual conclusion that this notion does not have what it takes to pick out a kind.

Before embarking on this route, however, it will be helpful briefly to compare the concept of consciousness-as-such with our concepts of specific conscious states, like pain, or seeing red, or tasting chocolate. I am interested here in such concepts as we might possess prior to any scientific investigations. Once we have arrived at scientific findings about either the determinable, consciousness-as-such, or its determinates, like pain, we might wish to incorporate these findings into augmented concepts of these properties. But before we can arrive at such scientific findings, we will need some everyday, pre-theoretical concepts. by which initially to pick out a subject matter for our scientific investigations.

In this connection, I would say that our pre-theoretical concepts of determinate conscious properties have a definiteness that is lacking from our concept of consciousness-as-such. In line with our earlier discussion, I take there to be two elements to such pre-theoretical everyday concepts. First, there are our phenomenal ways of thinking about determinate conscious states, our ways of thinking about those states by reactivating or reidentifying them. Second, I take there to be third-personal functional elements ("psychological" in David Chalmers' terms) in our everyday thinking about determinate conscious states. Some of the ways of referring to mental states that I earlier included included under the heading "third-personal concepts" will clearly be posterior to scientific investigation, for example, identifications in terms of physiology or sub-personal cognitive architecture. But prior to such discoveries we will already have some grasp, in everyday terms, of the functional roles played by determinate conscious states, as when we think of pain, say, as something caused by bodily damage and giving rise to avoidance behaviour.(7)

Now, I see no reason to suppose that our everyday concept of concsciousness-as-such contains either the phenomenal or functional elements characteristic of our everyday concepts of determinate conscious states. (8) Take first the question of whether we have a phenomenal concept of consciousness. It is difficult to know what to say here. Certainly we can imagine and recognize a list of determinate conscious states (pain, sadness, itches, and so on and on), and we can form some thought along the lines of "one of those". But whether this on its own amounts to any kind of concept, let alone a phenomenal concept akin to those we have for determinate conscious states, seems to me an open question. A list by itself does not tell us how to extrapolate to further cases. Because of this, I see no reason to regard the construction "one of those" as in itself doing anything to pick out conscious from non-conscious states.

The situation with our pre-theoretical functional grasp on consciousness-as-such seems more straightforward. We have almost no such grasp. Everyday thinking contains scarcely any ideas about what consciousness does. True, there is the idea that if a subject is "internally aware" of a state, then it is conscious, and (perhaps less strongly) that if any human state is conscious, then we will be "internally aware" of it. This fact, and how exactly to understand "internal awareness" in this context, will feature prominently in what follows. But beyond this connection with internal awareness, there is a suprising dearth of everyday ideas about any distinctive psychological role played by consciousness-as-such. We have no good a priori notion of any distinctive functional role played by all and only conscious states.

Testing Theories of Consciousness

Let me now switch tack, and ask instead how we test theories of consciousness. By asking this question, I hope to creep up on the concept of consciousness-as-such from behind. At least there seems to be an agreed methodology for testing theories of consciousness. By examining this methodology, we ought to be able to reconstruct the prior concept which identifies the subject matter of such theories. However, it will turn out that is hard to make good sense of the agreed methodology for testing theories of consciousness. This will in turn reflect adversely on the concept of consciousness-as-such.

While it is not often explicitly discussed, I take the standard procedure for testing theories of consciousness to be as follows. We humans look into ourselves, and check whether the states we are internally aware of coincide with the states in us that the theory identifies as conscious. This strategy seems appropriate across the board, from philosophical theories like Dretske's or Tye's intentionalism, through cognitive-functional theories like Baar's global workspace model, to physiological theories like Penrose's or Crick's and Koch's. We test the theory by seeing whether we can find states that we are internally aware of, but which don't fall under the theory's characterization of consciousness, or alternatively, whether there are states in us which do fall under the theory's characterization, but we aren't internally aware of. If there are states of either of these kinds, then this counts against the theory, while the absence of any such states counts in favour of the theory.

Let me illustrate this briefly by considering intentionalist theories Tye's or Dretske's, which equate being conscious with being a representational state of a certain kind. Emotions are a prima facie problem for such theories. For we are internally aware of our emotions, but they are not obviously representational. (What does anger represent,

or elation?) The standard counter is to argue that emotions are representational after all, despite first appearances. (Perhaps anger represents certain actions as unjust, and elation represents things in general as very good).

Intentionalist theories also face problems on the other side. For example, sub-personal representation (in early visual processing, say) is a prima facie problem, since we are not internally aware of such sub-personal states, even though they are representational. And to this the standard counter is to refine the theory, so as to be more specific about the kind of representation which is being equated with consciousness (perhaps it should enter into decisions in a certain way), and thereby to make sure that the theory does not attribute consciousness to any states we are not internally aware of.

The details do not matter here. My current concern is simply to draw your attention to the fact that theories of consciousness-as-such answer to the class of states we are internally aware of. Such a theories need to get the class of human states they identify as conscious to line up with the class of states we are internally aware of.

Now, you might wonder why I am belabouring this point. Isn't this the obvious way to test theories of consciousness? But I don't think it is obvious at all. On the contrary, I want to show you that there is something very puzzling here. It is not at all clear why a theory of consciousness should answer to the class of states we are internally aware of.

Internal Awareness and Phenomenal Concepts

As a first step, let us stop to ask what exactly is meant by "internal awareness" in this context. A natural answer is that we are internally aware of just that range of states for which we have phenomenal concepts. What shows that we aren't internally aware of sub-personal representations, for instance, or high blood pressure, to take another example, is that we cannot identify these states introspectively when we have them, nor can we recreate them in imagination.

Let me be clear here. My claim is not that a state's being conscious should be equated with with our having phenomenal concepts of that state. Whether this is so is a further issue, which will come into focus in a moment. My current claim is only about the notion of "internal awareness" I have been assuming when I have pointed out theories of consciousness answer to what we are "internally aware" of. I take it to be uncontentious that this notion at least can be equated with the availability of phenomenal concepts. To see this, suppose that there was some perceptual state, sensitivity to ultrasonic sound, say, which shared many of the features of paradigm conscious states (including guiding action and filling other higher cognitive functions), but for which we had no phenomenal concept whatsoever. That is, we were never able introspectively to identify it when it occurred, and we could never think about it by recreating it in imagination afterwards. It seems clear that a theory of consciousness which included such ultrasonic sensitivity among the class of conscious states would on this count be deemed to be deficient.

Perhaps my worry is now becoming obvious. Our methodological practice seems to rest on the assumption that the class of conscious human states coincides with the class of states for which we have phenomenal concepts. But, once we put this assumption on the table, it seems open to an obvious objection.

Let me pose this objection in terms which will be familiar, but which I have not used so far. A distinction is often made between sentience and self-consciousness. The standard assumption is that some animals, mice perhaps, and cats, are sentient, without being self-conscious. They have states that are like something, they have feelings, but they don't think about those states. In particular, they don't introspectively identify those states as mental states of a certain sort, nor do they think about them by recreating them in imagination. These further abilities require self-consciousness, which is only present in humans, and perhaps some higher primates. Self-consciousness requires concepts of conscious states, with which to think about conscious states, as well as just having conscious states.(9)

Now, if theories of consciousness are aiming to account for "what-it's likeness", as I have been assuming, then we need to understand them as aiming at the basic property of sentience, rather than the more sophisticated metarepresentational property of self-consciousness. But if this is right, then the standard methodology for testing theories of consciousness stands in need of further justification. For the availability of phenomenal concepts for certain states, while per se sufficient for the self-consciousness of those states, seems unnecessary for the sentience of those states. So now my worry is clear. If sentience is less than self-consciousness, as it seems to be, why should theories aiming at sentience be tested by seeing whether they correctly identify some category of states we are self-conscious of?

Sentience is Self-Consciousness

I can think of two possible answers to this question. The first assumes that consciousness requires higher-order thought, the second that internal awareness is a kind of observation. I shall consider these in turn.

The first answer in effect denies the distinction between self-consciousness and merely sentience, by arguing that it is a confusion to suppose that a state can be like something when it is not in some sense available to self-consciousness.

Note that, if this view is to help in the present context of argument, it needs to be upheld as an a priori thesis about our initial concept of consciousness, not as something we discovery empirically. We are currently trying to understand the puzzling logic by which theories of consciousness are standardly tested against the empirical facts. So it would fail to address this issue to postulate that the coincidence of sentient consciousness with self-consciousness emerges from this kind of empirical investigation. Rather, we need to suppose that, prior to any empirical investigation, conceptual analysis tells us that consciousness, in the sense of what-it's-likeness, just means some kind of internal awareness or self-consciousness.(10)

Putting to one side for the moment the plausibility of this a priori claim, note that the kind of "HOT" (higher-order thought) theory needed in the present context of argument has at least one virtue. Let us distinguish, following Peter Carruthers (forthcoming), between dispositional and actualist HOT theories of consciousness. Actualist HOT theories say that mental states are conscious only when they are actually the object of introspection. There are different versions of such actualist theories, depending on whether they conceive of introspection as more akin to thought, or as more akin to perception, but they share the feature that no particular mental occurrence is conscious unless it is actually being introspectively judged to be of a certain kind at that moment.

Dispositional HOT theories are not so restrictive about what it takes to be conscious. They allow that a given mental occurence is conscious if it can be the object of introspection, even if it is not currently being so introspected. So the dull pain in your left foot, or your visual perception of the car in front, are both conscious, on a dispositional HOT theory, although you aren't currently introspecting them, on the grounds that you are capable of such introspection, even if you aren't doing it now.

It is clearly the less aggressive dispositional version of a HOT theory that we need to make sense of the methodology by which we test empirical theories of consciousness. When we check to see whether an empirical theory of consciousness correctly identifies the states which we are "internally aware" of, we don't require it to pick out precisely those particular mental occurrences we are at some time actually internally aware of. Rather we want the theory to identify those types of mental states which we can be internally aware of, in the sense of having phenomenal concepts for states of that type. Thus, it would not be a problem for an empirical theory of consciousness that some of the particular occurrences it identifies as conscious do not happen to be introspectively identified or recreated in imaginative recall. All that is required is that those occurrences are of a kind which can so be objects of internal awareness.

Still, even if the kind of HOT theory currently needed is only a dispositional one, it still faces the obvious objection that its a priori standards for consciousness are unacceptably high.

To start with, note that standard HOT theories make consciousness a relational property of conscious states. For any standard HOT theory, determinate conscious states, such as being in pain, are conscious in virtue of the fact that their subjects are thinking about them, or at least could think about them. However, this then implies that just the same determinate states could occur (some organism could be in pain, say) but without any consciousness. Just keep everything else the same, but remove the higher-order thought. For example, consider a being like a cat, which I assume cannot think about its mental states. Now suppose that this cat instatiates the property which I think about when I am internally aware of a pain. A HOT theory will imply that the cat is indeed in pain, but that this pain is not conscious.

Of course, a variant style of HOT theory could avoid a commitment to unconscious pains (and emotions, sensory experiences, and so on), by being more restrictive about what is required for pains and other experiences. For example, they could say that the property of being in pain includes the feature that some first-order property is being thought about. The cat would then not be in pain (it only has a proto-pain, we might say), precisely because its state is not conscious, in that it does not incorporate any higher-order thought about anything.

This seems to me a more natural way to develop HOT theories. Rather than admit non-conscious pains and other experiences, it seems neater to continue to keep all experiences as determinate modes of consciousness, simply by including consciousness as a necessary condition of their presence. However, the underlying difficulty remains. Perhaps HOT theories can avoid unconscious pains and other experiences, simply by including higher-order thinking as part of what any such experiential state requires. But what they cannot avoid is the denial of consciousness to beings incapable of thinking higher-order thoughts. Whichever way we cut the cake, cats won't be conscious. Either they won't have pains at all, or they will have non-

conscious pains. And the same goes for all other beings who are incapable of thought about mental states, including one-year old human infants.

I take this to rule out HOT-style theories, at least as an a priori analysis of our concept of consciousness. Perhaps some form of HOT theory could emerge as the outcome of empirical investigation into consciousness, and indeed I shall return to this possibility later. But remember that in the present context of argument we are still trying to make sense of the logic of such empirical investigations, and so are trying to figure out what prior idea of consciousness sets the empirical agenda. And in this context it is surely unacceptable to claim that consciousness requires higher-order thought. Whatever content our prior concept of consciousness may have, it surely isn't such as to make it inconceivable that one-year old children should have conscious experience. It makes little sense to suppose that pure conceptual reflection could show us that it isn't like anything to be a baby that has hurt itself.

Inner Observation

Let me now turn to my second possible explanation for the standard methodology for testing theories of consciousness. On this explanation, the role of inner awareness is to observe a sample of conscious states, and so provide a data-base against which we can test empirical theories of consciousness. After all, if we consider other reductive theories in science, such as the reduction of water to H2O, or of temperature to molecular motion, they all rely on some pre-theoretical way of identifying instances of the target property, so as to provide a sample of cases against which to test the claim that all such instances possess the reducing property. So perhaps this is the role of internal awareness in testing theories of consciousness. Our internal awareness enables us to pick out, by directly observing their sentience, a sample of the sentient states that exist in the universe. And this then sets the stage for scientific investigation to identify some underlying property which constitutes the essence of these sentient states. (Having dismissed a priori HOT theories, I shall use "sentience" as a variant for "conscious" from now on.)

An initial set of queries about this observational model relates to the accuracy with which inner observation detects sentience. Other forms of observation are fallible in various ways. So it is natural to enquire whether inner observation of sentience is similarly fallible. More specifically, do we need to allow that this supposed faculty for observing sentience
(a) can make errors, in the sense of identifying some states as sentient when they are not, and
(b) be responsible for ignorance, in the sense of failing to register as sentient all human mental states which are sentient?

Now, questions of roughly this kind have been widely discussed in the philosophy of mind, under the heading of "self-knowledge". It is uncontentious that humans have some distinctive way of knowing about their own conscious mental states, and many philosophers have sought to explain how this works, and in particular to understand whether or not this faculty of self-knowledge is immune to mistakes.

Our current questions, however, are slightly different from the ones normally discussed under the heading of "self-knowledge", in that we are interested in internal judgements about which types of mental state are conscious, rather than internal judgements to the effect that a particular conscious state of some type is occurring ("I

am now in pain, tasting chocolate, or whatever"). That is, we are interested in whether internal awareness can go wrong in telling us positively that the property of being in pain, say, or seeing red, is conscious, and also in telling us negatively that the property of perceiving ultrasonically, say, or having high blood pressure, is not conscious. It wouldn't matter, from this perspective, if internal awareness went wrong now and then on particular mental states, classifying a particular sensation of cold as a pain, say, or failing to notice some particular conscious experiences at all. As long as it is right about which types of mental states are and are not conscious, it will provide accurate data against which to test reductive theories of consciousness.

Is internal awareness necessarily a good guide in this sense to which mental types are and are not conscious? Well, it seems hard to make sense of the idea that it could be prone to general errors of this kind, that is, that it could register instances of a certain mental type as conscious, while in fact they are not. (11) Perhaps the idea of inner awareness succumbing to ignorance is less problematic: maybe there should be room for certain kinds of perception, say, to count as sentient, even though we cannot introspectively identify or imaginatively recreate them.

However, I shall not pursue these issues any further here. This is because a high degree of accuracy is not essential to the observational model of the role of inner awareness in testing theories of consciousness. Suppose it were allowed that inner observation could indeed succumb to ignorance, and fail to identify certain sentient human states as conscious. Or suppose even (harder though this is to make sense of) that inner observation could fall into error, by presenting certain states as conscious when they aren't. Then the methodology for testing reductive theories of consciousness could simply be adjusted accordingly. If inner observation of sentience can be inaccurate, then to that extent a theory of consciousness is excused from lining up with its deliverances. (This would be in line with the standard scientific methodology for testing reductive scientific theories. While we need some pre-theoretical ability to identify water, or temperature, or whatever, to get the enterprise of reducing these kinds off the ground, these pretheoretic identifications are not regarded as inviolable. It is perfectly acceptable to allow a reductive theory to correct some of our pretheoretical judgements about what is and isn't water, if it is an otherwise attractive theory which fits the general run of cases.)

Is Consciousness a Kind?

My central worry about the observational model derives from different considerations. I doubt whether there is a kind, conscious, for inner awareness to detect. On the observational model, the role of inner awareness is to pick out those states which display the special property of consciousness. But I do not accept that there is any such special property.

I shall proceed by a kind of pincer movement. In this section I shall argue that there is no reason to suppose that any real kind is picked out by the description "the essence common to those states we are internally aware of". In the next section I shall then address the thought that, if inner awareness is a kind of observation, there must be such a kind, namely the kind observed by innner awareness. In response to this thought, I shall offer a non-observational account of the workings of inner awareness, which does not suppose that inner awareness detects some independently existing kind. That is, I shall show that inner awareness would work just as it does, even if

there were no division in nature between the states that are like something and those that are not.

So my first target is the idea of "the real essence common to those states we are internally aware of".

One initial problem here is whether we will find any interesting physical property (in the broad sense of "physical") which is peculiar to the states of which humans are internally aware. The range of states we are internally aware of is quite heterogenous. As well as pains, itches, tickles, and the various modes of sense experience, there are emotions, cogitations, and moods. There seems no obvious reason, on the face of it, why there should be any physical property common to this whole genus. Each species within the genus may share some common physical characteristic, but there may be no further physical property binding them all together which they do not share with all manner of bodily states.

However, let me put this worry to one side, and suppose that there are indeed one or more interesting physical properties peculiar to those states which humans are internally aware of. The next worry is then whether any of these will be either necessary or sufficient for consciousness in other kinds of beings. In humans, perhaps we are aware of all and only those states which involve 35-75 Hertz oscillations in the sensory cortex. But should we conclude on this basis that consciousness in any being will coincide with 35-75 Hertz oscillations in sensory cortices? The same problem will arise if we switch from physiological properties to functional ones. Maybe humans are aware of all and only those states which have a certain kind of representational role. But why infer on this basis that consciousness will everywhere coincide with this representational role?

At first sight it might seem that this is simply a standard inductive difficulty, which will arise whenever we try to identify the essence of some kind on the basis of some limited sample. However, this is not the real problem. It is true that the restriction to the human case does radically limit our sampling, and that this accentuates the standard inductive difficulty. But the problem goes deeper than that. The real obstacle is that we have no hold on what sort of property sentience is supposed to be, no clues about what sort of extrapolation from our sample is called for.

Here there is a contrast with kinds like water or temperature. In these cases, we don't just have some bare means of recognizing instances. We also have some a priori idea of what the kinds do. Water is wet, colourless, odourless, and so on. Temperature increases with inputs of heat, finds an equilibrium in closed systems, yields boiling and freezing at extreme values, and so on. Such prior ideas play a crucial role in guiding identifications of the scientific essence of such kinds. We don't just look for any physical property common to our limited finite samples of water or temperature. We look specifically for some physical property that will be characteristic of anything that is wet, colourless, and so on, or some physical quantity that will characterize any object's absorbtion of heat, will tend to equilibriate, will identify boiling and freezing points, and so on.

With consciousness it is different. As I pointed out earlier, when contrasting our notion of consciousness-as-such with our concepts of determinate conscious states, we have no real a priori idea of what consciousness-as-such does. Beyond the minimal assumption that consciousness registers in internal awareness, we don't have

a clue about what psychological role consciousness is supposed to play. The claim that a state is "conscious" tells us nothing about its psychological operations. So nothing guides the extrapolation of sample properties towards a unique target, some essence common to all conscious states. Without some prior notion of a role played by consciousness, nothing decides what sort of property should be sought.

Of course, there are plenty of functionalist "theories of consciousness", from representationalism to global workspace theories. However, given our lack of a priori ideas about the functional role played by consciousness, such theories can only be the outcomes of empirical investigation, arrived at by examining our introspected sample of conscious states, and identifying what they have physically in common. Yet it is precisely this kind of empirical investigation that is stymied, I am now arguing, by our lack of a priori ideas about the functional role of consciousness. Without some prior notion of what conscious states are generally supposed to do, the essence of consciousness cannot be fixed as one or other functional property which happens be common to the sample of states we are internally aware of.

The observational model of internal awareness offers the possibility of identifying the reference of "conscious" as that essence which scientific investigation will reveal to be characteristic of the range of states we are internally aware of. But it turns out that scientific investigation does not have the wherewithal to complete this job. Without some prior specification of the psychological role consciousness is supposed play, the enterprise of identifying its scientific essence cannot get off the ground. And so the idea that "conscious" refers to that essence turns out to be empty.

Perhaps this is too quick. Isn't there a case for bringing back HOT-style theories at this point? Even if such theories cannot be vindicated a priori, they certainly identify a salient broadly physical characteristic peculiar to the states we are internally aware of. After all, the one such feature which is uncontroversially peculiar to those states is that we can reidentify them introspectively and recreate them in imagination. So why not accept this as the underlying attribute which empirical investigation reveals to be essential to consciousness?

This suggestion has attractions, but it seems to me open to an objection based on the point made earlier against a priori HOT-style theories. There I urged that it is surely a conceptual possibility at least that beings without higher-order thought, like human babies, may nevertheless have conscious states. So, if scientific investigation is going to show that this conceptual possibility is not actual, this will presumably be on the basis of positive empirical evidence that non-reflective beings are not in fact conscious. However, the proposed route to an a posteriori HOT theory does not really invoke any such positive evidence. It is entirely predetermined by the methodology that the most obvious characteristic common to our sample of "observed" states will be that we are internally aware of them. Whatever scientific investigation may or may not discover about these states, we know beforehand that they will all be states that we can introspectively reidentify and imaginatively recreate, for this is simply what we are assuming is needed to "observe" the consciousness of those states. Given this, it would surely be wrong to regard this fact alone as ruling out what we have already agreed to be a live conceptual possibility, namely, the possibility that something less than internal awareness may be required for consciousness itself, as opposed to the observation of consciousness.

A Different Model of Internal Awareness

Some of you may feel that the conclusion being urged in the last section comes close to a reduction of my overall stance. Surely I have followed the argument too far, if I am in danger of concluding that those states that are like something do not constitute a kind. Isn't it obvious that there must be such a kind? Something must have gone wrong with my argument, you may feel, if it forces us to deny this. Indeed perhaps my conclusion reflects back on my initial physicalism. If the cost of physicalism is to deny the realty of consciousness, then surely we ought to take another look at dualism.

But let me ask you to bear with me for a moment. Perhaps our conviction that there must be a kind here derives from too ready an acceptance of the observational model, and on its accompanying idea that consciousness is the kind that is there observed. For note that, once we start thinking of inner awareness on the model of observation, then this accompanying idea becomes irresistible. As soon as we embrace the observational model, we are forced to think of inner awareness as a kind of scanner, searching through the nooks and crannies of our minds, looking for those states with that special extra spark, the feature of consciousness. The observational model thus automatically carries with it the idea of some kind to which the observational faculty responds.(12)

However, there is another way of thinking about inner awareness, which accounts happily for all the facts to the hand, but which doesn't trade on the idea of inner observation, and so a fortiori doesn't commit us to the existence of a property which triggers such observations.

Suppose that there isn't anything special about the states which we are internally aware of (the states which we know to be "like something"), apart from the fact that they are hooked up to higher-order thinking as they are, via introspection and imagination. That is, the only thing that is special about them is that we have phenomenal concepts for them. We can introspectively reidentify and imaginatively recreate them. On this suggestion, then, we are internally aware of certain states simply because they enter into our thinking in a special way. They don't enter into our thinking in this special way because they have some further special nature.

Let me draw out a natural, if somewhat striking, corrolary. Any state that was similarly hooked up to our thinking would thereby we one which we knew what it was like to have. We are internally aware of certain states because they are used in our thinking, in a way that other states aren't. We think with them, when we deploy phenomenal concepts of them. This is what makes us pick them out as states that are "like something". But don't need to suppose that there is anything else special about such states. Any state that was similarly used in our thinking would strike us as a state which is like something. Once a state is so used, its being gets into our thinking, so to speak, and its nature ("what it is like") thus becomes part of our mental life, in a way that is not true for other states of the world.

Let me use an analogy. To think of inner awareness as observation of a special kind is like thinking of television as a medium which identifies a special kind of "televisualisable" event. Imagine an innocent who thought that some worldly events display a special glow, a kind of luminance, and that television cameras were instruments somehow designed to detect this glow and transmit events with it to our screens. This would be a mistake. Even if some events are more suitable for televising

than others, for aesthetic or commercial reasons, this plays no part in the explanation of television transmissions. Any event can appear on your set, once it is appropriately linked to it via televion cameras and transmission stations. Similarly, I say, with consciousness. Any event is capable of registering in internally awareness, once it is linked to it via introspection and imagination.

Conclusions

I have now argued against both possible explanations of the standard methodology for testing empirical theories of consciousness. The a priori equation of sentience with self-consciouness is unacceptable. And the idea of inner awareness as observation both fails to pick out a kind and is in any case unnecessary to account for the workings of inner awareness.

From the perspective we have now reached, we can see what goes wrong with attempts to construct scientific "theories of consciousness". There is indeed one unproblematic distinction associated with the methodology used to construct such theories. But theories of consciousness want to get beyond this distinction to some further boundary in reality, and unfortunately there is nothing there for them to find.

The unproblematic distinction is between states which are thought about phenomenally, and those which are not. As I showed in earlier sections, a satisfactory physicalist account of this distinction is straightforward. Phenomenal thought is simply a matter of imaginative recreation and introspective reidentification. There seems no barrier to an understanding of how brains can generate recreations and reidentifications. So there is no difficulty for a physicalist about the difference between states which are objects of phenomenal thought and others.

Dispositional higher-order thought theories of consciousness want to equate the distinction between conscious and non-conscious states with the unproblematic difference between states that are objects of phenomenal thought and others. However, this is not faithful to our initial pre-theoretical concept of consciousness. There may not be very much to this pre-theoretical concept, as I have argued throughout, but it does at least resist the a priori identification of consciousness with higher-order thought. It is conceptual possible that there should be states which are conscious though not self-conscious.

Unfortunately, the pre-theoretical concept of consciousness then casts us adrift. Having pushed us away from the identification of consciousness with self-consciousness, it fails to offer any other hold on where the line between consciousness and non-consciousness might fall. The concept of consciousness-as-such thus turns out to be an irredeemably vague concept, whose application to states other than the phenomenally self-conscious is quite indeterminate.

Can we rest here? Not happily. The notion of consciousness may be too thin to pick out any real kind, and so fail to point to anything beyond the category of phenomenal self-consciousness. But there remain serious further issues which are standardly taken to be bound up with questions of consciousness, and these issues seem to demand answers.

I am thinking here primarily of moral questions about the treatment of non-human creatures. What moral principles should govern our conduct towards cows, or fish, or

lobsters, or indeed possible intelligent machines or extra-terrestrial organisms? It seems natural to suppose that answers here will hinge on whether such creatures are conscious. It would be all right to drop live lobsters into boiling water, for example, if they feel no conscious pain, but not if they do.

Perhaps some progress here can be made by switching back from the determinable, consciousness-as-such, to more specific determinates like pain. That is, we might be able to establish that cows and lobsters, say, should be objects of moral concern, even in the absence of a general theory of consciousness-as-such, simply by establishing that they have pains. (13)

One issue which arises here is whether even our concepts of such determinate states as pain are contentful enough to pick out a real kind, and so determine which creatures are in pain. This is really the topic for another paper. Let me simply make two comments. First, and in line with my discussion in earlier sections, I take it that our pre-theoretical concepts of determinate mental states are at least more contentful than our concept of consciousness-as-such. Second, I suspect than even these concepts will be infected by some degree of vagueness, with the result that it is to some extent indeterminate which real kinds they refer to. Given these two points, my guess is that cows can be shown definitely to experience pain, but that it may be indeterminate whether lobsters do, and even more indeterminate with exotic creatures from other parts of the universe and the technological future.

In any case, the strategy of trying to resolve the moral issues by switching from the determinable, consciousness-as-such, to determinables, like pain, can only have limited application. For it will only yield a way of deciding issues of consciousness in connection with those determinate conscious states for which we happen to possess the appropriate concepts. Yet it seems clear that there can be determinates of the determinable, consciousness-as-such, which we humans are unable to conceptualise phenomenally. The experience of bats when they echolocate, for example, is arguably a determinate conscious property which we humans are unable to conceptualise in phenomenal terms. And there would seem no limit in principle to the humanly unconceptualised modes of consciousness that could be present in intelligent machines or extra-terrestrial creatures.

I am not at all sure how to make further progress at this stage. An optimistic view would be that further scientific knowledge will help us decide such tricky questions as whether it is morally permissible to immerse live lobsters in boiling water. However, even if it can, I do not think that it will do so by showing us whether or nor lobsters are conscious. If the arguments of this paper are any good, the concept of consciousness is too vague to fix an answer to that question, however much scientific knowledge becomes available.

Footnotes

(1) When I talk about a mental "state", I shall simply mean the insatiation of a mental property. I shall take the particulars which insatiate such properties to be ordinary organisms, like people or possibly cats. This means that my basic notion of consciousness is that of an organism being conscious at a given time. For those readers who prefer to focus on "state consciousness", rather than "creature

consciousness", let me observe that, when an organism is conscious on my account, this will be in virtue of its being in one or more mental states which are determinates of the determinable, being conscious. These determinate mental states are thus naturally counted as "conscious" states. (Compare: my car is coloured in virtue of being red, or grenn, or whatever: red and green and so on are thus colours). As for "consciousness of", the other notion whcih is sometimes taken to be basic, I do not assume that all conscious states have intentional objects (though some certainly do), nor that all intentional states are conscious.

(2) Note that Mary will be able to make such reidentifications even if she doesn't have a word for the experience. Nor need she yet be able to identify the property she can now think about in phenomenal terms with any of the properties she could previously think about third-personally.

(3) Nor, a fortiori, does the physicalist story rest anything on the dubious idea of direct acquaintance with such phenomenal properties. I find the anti-physicalist story especially puzzling at this point. In particular, how is the change in Mary (her now "knowing what seing red is like") supposed to be sustained after she stops having her new experience? Can she now recall the phenomenal property to her mind at will, so as to reacquaint herself with it? Or can her memory reach back through time to keep acquainting her with the earlier instance? Both these ideas seem odd, but something along these lines seems to be needed to explain why Mary continues to "know what the experience is like" after the experience is over, if such knowing depends on acquaintance with a phenomenal property.

(4) What is the status of the claim that you can only know what an experience is like once you have had it yourself? It does seem, give or take a bit, that actual human beings cannot imaginatively recreate or introspectively reidentify any conscious experiences that yhey have not previously unergone. But note that the explanantion I have given for this phenomenon implies it to be a pretty contingent matter. There seems nothing impossible about creatures being born with imaginative and introspective abilities, prior to any experience. (Simply imagine that the "moulds" or "templates" involved, and that dispositions to use them, are "hard-wired", that is, grow independently of any specific experiences). There is also an interesting question here about the link between imaginatively recreation and introspective reidentification as such, whether or not these powers derive from prior experiences. Again, these seem to go hand in hand in actual human beings (though here the explanation is not so obvious), while once more there seems nothing impossible about creatures in whom they would come apart.

(5) What does make it the case that phenomenal concepts refer as they do? This is a further question, on which I shall not say much (though see footnote 8 below). In a moment I shall argue that, when phenomenal concepts are deployed in imagination or introspection, their employment phenomenally resembles the experiences they refer to, and moreover that this resemblance is importanmt for understanding the mind-brain issue. But I should like to make it clear that I certainly do not take this resemblance to contribute in any way to the referential power of phenomenal concepts. (I should alos like to make it clear that I do not take the representation of experiences to be the primary function of imagination. It seems clear that the ability to visually imagine a tiger, say, would in evolutionary terms first have been useful because it enabled us

better to think about tigers. Only later would this ability have been co-opted to enable us to think about experiences themselves).

(6) Some readers might not be persuaded that imaginative recreations of experience really feel like the experiences themselves. Does an imagined pain really feel like a real pain? I myself do think there is a real phenomenal resemblance, especially with pains and colour experiences, and that this is part of what seduces people into the antipatheitic fallacy. But I do not need to insist on this. Fot I can restrict my diagnosis of anti-physicalist intuitions to the other kind of use of phenomenal concepts, namely, in thoughts which deploy introspective identifications, rather than imaginative recreations. These uses unquestionably feel like the experiences themselves, since they use actual experiences, brough under some categorisation, and not just some imagined recreation. I would like to stress this point, given that my previous explanations of the "antipathetic fallacy" have focused on imaginative recreations, rather than introspective identifications, for reasons which now escape me. It is perhaps also worth stressing that the points made about phenomenal concepts in earlier sections, and in particular my explanation of how they help physicalists to deal with Jackson and Kripke, are quite independent of my account of the antipathetic fallacy.

(7) Note that it would be a mistake to infer, from the fact that a concept is available prior to active scientific investigation, that it must be unconnected with any testable empirical theory. It may well derive its content from some empirical "folk" theory. (Note also that in that case, as with all "theoretical concepts", it won't be the concept's existence, so to speak, but its satisfaction, that depends on the truth of empirical theory. I can use "pain" to express "that property, if any, which results from damage and causes avoidance behaviour" even if it is empirically false that there is any such property. If there is no such property, then my term will still be meaningful, but will simply fail to refer to anything).

(8) Should we think of our everyday word "pain" as ambiguous, equivocally expressing separate phenomenal and fuctional concepts? I am not sure. An alternative would be to understand "pain" as referring to that property, if any, which satisfies both concepts. (Note that, on this option, "pain" would arguably end up non-referring if epiphenomenalism, rather than physicalsim ir interactionism, were true). A further question which arises at this point is whether "pure" phenomenal concepts mustbe augmented by at least some functional specifications if they are to have any power to refer determinately.

(9) Let this be a stipulation of what I mean by "self-conscious". There is also a weaker notion, which would be satisfied by creatures (some apes perhaps, or indeed people with "theory of mind" deficiencies) who can think of themselves as one creature among others, but do not have any thoughts about mental states.

(10) The role of empirical investigation would then be to tell us which states are conscious (that is, objects of higher-order thought) in which creatures. This would indeed make good sense of much psychological research. Information about 35-75 Hertz oscillations in the human sensory coretx, say, seem of no great relevance to our understanding of consciousness in any possible creature. But it could help identify those states which are objects of higher-order thought in human beings.

(11) Someone could of course make the theoretical mistake of holding that believings, say, are conscious, when they aren't. But this may be just because they haven't introspected carefully enough. What does not seem possible is that inner awareness could itself mislead us on such a matter, by taking episodes of a certain type to be conscious, when they are not.

(12) Note also how any residual dualist inclinations will strongly encourage this observational package. On the dualist picture, conscious states are peculiar in involving phenomenal aspects of reality, and it is then very natural to think of inner awareness as sensitive precisely to this property of phenomenality.

(13) We can take it at this stage that pains carry consciousness with them. The idea of non-conscious pains was an artefact of HOT theories' a priori equation of consciousness with self-conciousness, and we have rejected such theories precisely because this equation is mistaken.

References

Armstrong, D., 1968, A Materialist Theory of the Mind, London, Routledge and Kegan Paul.

Baars, B., 1988, A Cognitive Theory of Consciousness, Cambridge, Cambridge University Press.

Carruthers, P., forthcoming, Phenomenal Consciousness.

Chalmers, D., 1996, The Conscious Mind, Oxford, Oxford University Press.

Churchland, P. and Churchland, P., 1998, On the Contrary, Cambridge, Mass., MIT Press.

Crick, F., 1994, The Astonishing Hypothesis, London, Simon and Schuster.

Dennett, D., 1991, Consciousness Explained, London, Allen Lane.

Dretske, F., 1995, Naturalizing the Mind, Cambridge, Mass., MIT Press.

Jackson, F., 1986, "What Mary Didn't Know," Journal of Philosophy, 83.

Lycan, W., 1996, Consciousness and Experience, Cambridge, Mass., MIT Press.

McGinn, C., 1991, The Problem of Consciousness, Oxford, Basil Blackwell.

Papineau, D., 1998, "Mind the Gap", in J. Tomberlin, ed, Philosophical Perspectives, 12.

Papineau, D, 2000, "The Rise of Physicalism", in B. Loewer, ed., Physicalism and its Discontents, Cambridge, Cambridge University Press, and in M. Stone and J. Woolf, eds, The Proper Ambition of Science, London, Routledge.

Penrose, R., 1994, Shadows of the Mind, Oxford, Oxford University Press.

Rosenthal, D, 1996, "A Theory of Consciousness", in N. Block, O. Flanagan and G. Güzeldere. eds, The Nature of Consciousness, Cambridge, Mass., MIT Press.

Shallice, T., 1988, From Neuropsychology to Mental Structure, Cambridge, Cambridge University Press.

Tye, M., 1995, Ten Problems of Consciousness, Cambridge, Mass., MIT Press.