What follows are the drafts of chapters 2 through 6 of a book on consciousness, provisionally entitled "The Phenomenon of Consciousness".

The material below is reasonably self-contained. As to the rest of the book, Chapter 1 presents the initial reasons for favouring materialism about consciousness, while two final chapters consider problems that arise for detailed scientific theories of conscious states.

The five chapters below are:
Chapter 2: Jackson's Knowledge Argument and Conceptual Dualism
Chapter 3: Kripke's Modal Argument and Direct Phenomenal Reference
Chapter 4: Phenomenal Concepts
Chapter 5: The Explanatory Gap
Chapter 6: The Intuition of Distinctness

If this is a bit much, I suggest skipping the first half of chapter 3 (until section 3.5) and the last two-thirds of chapter 5 (from 5.3 on). These parts rehearse relatively familiar material about Kripke's argument and the explanatory gap.

Much of chapter 2 is also pretty standard material on Jackson, but I do use this chapter to set up some distinctive theses.
The Phenomenon of Consciousness
David Papineau

Chapter 2

Jackson's Knowledge Argument and Conceptual Dualism

2.1 Introduction

The last chapter offered an argument for a materialist view of consciousness, where materialism is to be understood as a matter of property identity. Conscious properties are identical to material properties—that is, either strictly physical properties, or physically realized higher-order properties.

Still, while I am a materialist about 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 conscious properties. Moreover, I think that is crucially important for materialists to realize that conscious properties can be referred to in these two different ways. Materialists who do not acknowledge this, and there are some, will find themselves unable to answer some standard anti-materialist challenges.

I shall call these two kinds of concepts "phenomenal" concepts and "material" concepts. I shall have plenty to say about both kinds of concepts in what follows. But it will be helpful to start with a rough initial characterization.

Material concepts are those which pick out conscious properties as items in the third-personal, causal world. Most commonly, these will be role concepts, associated with some specification of a causal role, such as pain's role in mediating between bodily damage and avoidance behaviour. But I want also to include under this heading directly physical concepts which identify their referents in terms of some intrinsic physical constitution.

Opposed to such material concepts are phenomenal concepts. This category is less familiar. The general idea is that, when we use phenomenal concepts, we think of mental properties, not as items in the material 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 materially, in terms of nerve messages, brain activity, involuntary

---

1 In Remnants of Meaning (1987) Stephen Schiffer combines a "sentential dualism" with an ontological physicalism about propositional attitudes. There are affinities between this and my "conceptual dualism" about conscious experiences, though also many specific differences.

2 Role concepts of properties can be of two types. They can name whichever first-order property realizes the role, or they can name the higher-order property which constitutes the role. This distinction won't matter in the present chapter. It is perhaps worth noting that it is not a priori, even if it is true, that material concepts of this role kind should name material properties as defined in the last chapter; for it is not a priori, even if true, that the relevant roles will be physically realized.

I could perhaps also have included some perceptual concepts under the heading of material concepts of conscious states, such as visual concepts of certain brain states. But since such concepts of brain states play no prominent role in my arguments until chapter 6, it will simplify things to leave them out.
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.  

Now, as a materialist, I hold that even phenomenal concepts refer to material properties. In distinguishing phenomenal concepts from material concepts, I do not wish to suggest that they refer to different entities. The argument of the last chapter still give us reason to take the two kinds of concepts to make common reference to material properties. The idea, then, is that we have quite different ways of thinking about pain, say, or tasting chocolate, or seeing an elephant, both of which refer to the same material properties in reality. By way of obvious analogy, consider the case where we have two names, "Tony Curtis" and "Bernie Schwartz", say, both of which refer to the same real person.

We might say that the difference between phenomenal and material concepts is a difference at the level of sense, not reference. As in standard cases of co-reference, we have two terms which refer to the same entity, but in different ways—that is, in virtue of different senses. There will be many questions to answer about these distinct modes of reference, and in particular about the mode in which phenomenal concepts refer. But the underlying assumption will remain that these different modes both point to the same objective material property.

If phenomenal and material concepts are quite distinct at the level of sense, there will be no a priori route to the identification of their referents. Examinations of the concepts themselves will not tell us that they refer to the same properties. Such knowledge can only be arrived at a posteriori, on the basis of empirial evidence about their actual referents. Still, this will not worry materialists who defend materialism in the way outlined in the last chapter. For nothing in that line of argument depended on any a priori analysis of concepts.

Ned Block (forthcoming) has recently coined some useful terminology. He uses the terms "inflationists" for philosophers who recognize the existence of phenomenal concepts. Not all materialists are inflationists. As we shall see, a number of leading materialist philosophers, including David Lewis and Daniel Dennett, deny the existence of phenomenal concepts, and hold that all references to conscious states are made using material concepts alone. Since these philosophers do not recognize any distinctive conceptual apparatus for referring to conscious states, Block calls them "deflationists".

David Chalmers (forthcoming) has also devised a terminology to mark this distinction among materialists. He uses "type A" to refer to those materialists—deflationists, in Block's terms—who do not recognize distinct phenomenal concepts, and take all reference to conscious properties to proceed via material concepts. "Type B"

---

3 My distinction between "phenomenal" and "material" concepts is similar to David Chalmers' distinction between "phenomenal" and "psychological" concepts (Chalmers, 1996). But his "psychological" concepts are specifically role concepts, and for present purposes it is more convenient to work with my more general category of non-phenomenal "material" concepts. In chapter 4, however, I shall have occasion to make use of Chalmers' "psychological" category.
materialists are then those inflationists who hold that we possess special phenomenal concepts of conscious states, in addition to any material concepts.\footnote{To forestall one possible confusion, let me make it clear that I do not take my conceptual "dualism" itself to demand any special non-material ontology. In my view, the deployment of phenomenal concepts depends on material processes in thinkers' brains, just as much as the deployment of any other concepts. Indeed, I shall shortly say something more about the brain processes that might underly the deployment of phenomenal concepts. It is true that the ontology of concepts themselves is a somewhat obscure matter. Some philosophers would argue that they are a species of abstract entity, akin to numbers. I have my doubts about this, and would hope to parse away any such references to concepts as abstract objects. But, in any case, it is worth observing that, even if concepts were abstract objects, this kind of non-materiality would not be peculiar to phenomenal concepts, but would be common to all concepts.}

2.2 Jackson's Knowledge Argument

The best way to demonstrate the existence of phenomenal concepts is via Frank Jackson's "knowledge argument" (1982, 1986). Jackson himself takes this argument to demonstrate the existence of distinctive phenomenal properties, that is, conscious properties which cannot be identified with any material properties, and which therefore refute materialism. I think that his story does not establish this anti-materialist conclusion, and will shortly argue as much. But at the same time it does provide an excellent way of establishing the existence of distinctive phenomenal concepts.

Jackson's argument is made graphic by his well-known "Mary" thought-experiment. Mary is some future cognitive scientist. She is an absolute authority on human vision, and in particular on colour perception. She has complete material knowledge about what goes on in humans when they see colours. She knows all about light waves, and reflectance profiles, and rods and cones, and about the many areas concerned with vision in the occipital lobe, and what they each do, and about the kinds of circumstances that produce different colour experiences, and the kinds of illumination that produce colour illusions, and so on, and on.

However, apart from this, Mary has had a somewhat unusual upbringing. She has never seen any colours herself. She has lived all her life inside a house painted black and white and shades of grey. All her knowledge of colour vision is book learnin', and none of her books contains any colour illustrations. She has a TV, but it is an old black-and-white set.

Then one day Mary walks out of her front door, and sees a red rose. At this point, Jackson observes, Mary learns something new, something she didn't know before. She learns what it is like to see something red.

Jackson takes this to show that Mary becomes acquainted with some new property of red experiences, the "conscious feel" of red experience. After all, before she came out of the house she already knew about every material property of red experiences. If she learns about something new, argues Jackson, this must involve her now knowing about some further feature of red experiences, the conscious feature, which cannot therefore be identical with anything material.
However, materialists who recognize phenomenal concepts needn't accept this argument. They can respond that, while there is indeed a genuine before-after difference in Mary, this is just a matter of her thinking in new ways, and in particular of her acquiring a new concept of seeing red. There are no new properties in the offing. The property she refers to with this concept is still a perfectly good material property, that material 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 material concepts, even before she saw the rose.

2.3 Denying the Difference

Let me go a little more slowly. Not all materialist philosophers respond to Jackson's argument in this way—that is, by arguing that Mary is changed at the level of concepts, even if not by any acquaintance with new phenomenal properties. I shall consider two alternative materialist responses which do not even concede this much of a before-after change in Mary. They deny that she so much as acquires any new concepts. These are "deflationist" responses to the Jackson's argument, in that they see no reason to credit Mary with anything but material concepts, even after she leaves her house. Exposing the deficiencies in these deflationist strategies will help to make it clear why materialists are well advised to recognize distinctively phenomenal concepts.

First, there are materialist philosophers, most prominently Daniel Dennett, who aim to stop the Mary argument before it starts, by denying that Mary displays any significant before-after difference in the first place. Of course, there is one trivial before-after difference, which can be agreed on all sides. This is that Mary has a new experience after she comes out of the house, an experience of a kind she has never had before. This is not at issue, for there is nothing in this to provide any argument against materialism. Materialists are just as well placed as anybody else to explain this difference. Materialists think that conscious experiences are identical with certain material occurrences in the brain. So materialists can simply say that this before-after difference in Mary, that she has now had an experience she hadn't had before, is simply that certain material states, namely, those which constitute red experiences, have now occurred in her, when before they hadn't.

The more important question is whether there are any further before-after differences in Mary, consequent on her having had this experience. Jackson wants to say that, in addition to having had the experience, she now also knows something she didn't know before, namely, what the experience is like. This knowledge isn't just a matter of once having had the experience itself. It is something that remains with Mary after the experience is over. With luck, she'll now retain her knowledge of what seeing red is like throughout her life. It is this further change that Jackson takes to present a problem for materialists.

Daniel Dennett (1991) denies there would be any such before-after change in Mary. He points out that Jackson's thought-experiment credits Mary with an awful lot of knowledge even before she comes out of the house. After all, she is supposed to know everything that can be known about colour vision using non-phenomenal concepts. In addition to her physiology, optics, and computational neuroscience, this will include a
vast body of everyday knowledge about the characteristic causes and effects of red experiences.

For normal people, of course, it is first-hand acquaintance that yields much of this non-phenomenal knowledge of colour experience. It is through having red colour experiences that most of us find out, for example, that roses and ripe tomatoes are red, that red experiences are associated with other psychological states, such as sensitivity to danger, and so on. So, for most of us, there is a real before-after difference in knowledge occasioned by new experiences. The experiences indeed give us information we didn't have before.

Still, argues Dennett, we shouldn't allow this intuitive truth, which rests on normal facts about normal people, to be "pumped" by the Mary thought-experiment into the conclusion that experience would give new knowledge even to someone who already knew everything that could be known in material terms about red experiences. Such a person—a Mary—would already know all the things that ordinary people find out from first hand experience. So she, unlike ordinary people, wouldn't gain any new knowledge from her experience. True, Dennett allows, it is difficult for us to conceive clearly of someone who really does have total material knowledge about red experiences, so far removed from reality are the conditions of the thought-experiment. Because of this, we naturally assimilate Mary to the normal case of people who do learn from first-hand experience. But if we attend carefully to the specifications of the Mary thought-experiment, we should resist this assimilation.

Or so at least Dennett argues. However, I find this line of argument quite implausible. Dennett is looking in the wrong place for the relevant before-after differences. The important changes occasioned in Mary, and signalled by the phrase "she now knows what it is like", are nothing to do with her gaining any new material knowledge, such as knowledge of which things cause red experiences, or of which other materially characterised mental states characteristically accompany red experiences. By hypothesis, she already has all this knowledge. Rather, the important changes in Mary are that she acquires some quite new powers of thought, of a kind she simply didn't have before.

In particular, once she has seen something red, Mary will be able to recreate this experience in imagination. Moreover, she will be able classify new experiences introspectively as of the same kind. The most natural way of reading the expression "to know what something is like" is as referring to these two new powers. Mary is changed, not in getting more knowledge of the material kind she previously had, but in acquiring these powers of imagination and introspection.

Thus, someone who undergoes a new kind of experience will later be able to imagine what the experience is like, in a way they couldn't before. They will have a grasp of the redness of red experience, so to speak. In addition, someone who undergoes a new kind of experience will later be able introspectively to tell whether further experiences also feel like that. They will be able directly and introspectively to classify new experiences as having that characteristic redness.

Now, the analysis of these imaginative and introspective powers will occupy much the rest of this book. But at this stage, even before we go into the details, we can see why
Dennett's line is quite unconvincing. Dennett is committed to the view that the pre-experiential Mary, in virtue of her encyclopedic knowledge, can already imagine what it is like to see something red, and can already classify further experiences directly and introspectively as of that type.

This is surely quite implausible. In the next two sections I shall offer a natural explanation of why Mary can't do these things, prior to her own red experience. But, even prior to this explanation, it seems clear that she won't be able to do these things before she emerges from her house. No amount of book learning will tell her how to create the experience of red in imagination, or how introspectively to classify experiences as of that type. To suppose otherwise is to suppose that book learning on its own will somehow show you what it feels like to have a red experience, and show you how to judge introspectively whether or not some further experience involves that feeling. This seems wacky.

For any readers who may remain unconvinced, I need not press the point at this stage. The rest of this chapter will make it amply clear why Dennett's line is both unnatural and unnecessary for a materialist. The reason Dennett himself takes this line, I suspect, is that he is strongly committed to some kind of "deflationist" analysis of concepts of mental states. He assumes that there is no other respectable way of thinking about mental states apart from thinking of them in terms of roles, that is, as states with certain canonical links to behaviour and perhaps other similarly identified mental states. And of course, if you do take this to be the only respectable way to think about mental states, then you must conclude that Mary's new experience couldn't possibly lead to any new information, since she already had all the information that could possibly be framed using role concepts of mental states.

Still, it seems desperate to end up denying, as does Dennett, that there is no real before-after difference in Mary. If that is the cost of deflationism, then so much the worse for deflationism.

As it happens, it is by no means clear that deflationism as such requires Dennett's desperate measure. In section 2.6 below I shall consider a second version of deflationism, the "ability hypothesis", that admits a real before-after difference in Mary, of just the kind I have been insisting on, yet which denies that this difference forces us to inflationist conceptual dualism.

However, before we consider this "ability hypothesis", we need to look more carefully at the supposed before-after differences in Mary. Now that we are agreeing, contra Dennett, that there are real before-after differences, we had better make sure that they do not imply some new acquaintance with phenomenal properties, in the way Jackson supposes.

I said that the crucial differences lie in Mary's new powers of imaginative recreation and introspective classification. In the next two sections I shall accordingly consider these in turn, showing in each case that there is no legitimate argument from either before-after difference to distinct phenomenal properties. The Mary argument does not establish any dualism of properties.
Then I shall return to the "ability hypothesis". The issue here is whether the before-after change in Mary at least implies a dualism of concepts, even if not of properties. The sophisticated deflationism of the "ability hypothesis" aims to argue that the before-after differences in Mary do not imply even this much. I shall argue against this that the "ability hypothesis" is inadequate, and that the story of Mary does demonstrate the existence of distinctive phenomenal concepts.

1.4 Imaginative Recreation

The first before-and-after change to be considered concerns Mary's new powers of imaginative recreation. Once she has seen red, Mary can recreate the experience of seeing red, whereas before she couldn't. Mary could of course always imagine, in the third-person, so to speak, that somebody else was seeing red, in the sense that she could posit such-and-such material occurrences in that person. But now she has a new ability. She is able to imagine having the experience itself, from the inside, as it were. She can now relive the experience, as opposed to just thinking about it.

Anti-materialists want to account for this change in terms of Mary's new acquaintance with some non-material property. The anti-materialist story would go something like this. When Mary experiences red, she becomes acquainted with the characteristic phenomenal feature of red experiences. And henceforth this acquaintance enables her to imagine the experience in question, since she will now be able to call this property to mind, and thereby recreate in her mind the characteristic phenomenal feel of red experiences.

However, there is an obvious alternative materialist story to be told. This accounts equally well for the fact that you can't imagine an experience prior to having it, and does so without invoking any special phenomenal properties. 5

Here is the obvious materialist 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 the parts of the brain required for the imaginative recreation of some type of experience, if some actual experience of that type has previously activated those parts.

There is now plenty of evidence to support this hypothesis about imaginative recreation. Data from brain scans and similar techniques show directly that imagination activates some of the same parts of our brains as are activated by actual experiences of the relevant type. Moreover, studies of patients with brain lesions

5 Nor, a fortiori, does the materialist story rest anything on the dubious idea of direct acquaintance with such phenomenal properties. I find the anti-materialist story especially puzzling at this point. In particular, how is the before-after change in Mary supposed to be sustained after she stops having her new experience? Can she now reproduce the phenomenal property in 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 is permanently changed by her experience, given that the change depends on her direct acquaintance with a phenomenal property.
shows that damage to the relevant areas will also destroy imaginative abilities. Someone with damage to the relevant parts of the visual cortex will not only lose the ability to see colours, but also the ability to imagine them. Both these lines of evidence strongly suggest that imaginative recreation is a matter of "turning on" some characteristic pattern of brain activity that was first created by an original experience.

These remarks of course only gesture at a complex body of empirical data. But they suffice to indicate how materialists might explain Mary's new power of imaginative recreation, yet deny that it demands any new non-material property. Mary's new power does not depend on any acquaintance with such a phenomenal property. Rather, her brain is lastingly altered in certain ways, and this now allows her imaginatively to recreate an experience that she could previously only think about in a materially. Seen in this way, it is clear there is nothing in the idea of imaginative recreation to worry materialists.

1.5 Introspective Classification

The other change in Mary was to do with introspective classification. Once she has seen red, she can introspectively classify further experiences as of that type. We can think of Mary as acquiring a new classificatory category, for which she might have no word, but which she can apply to particular new experiences.

Anti-materialists will again maintain that this new power of introspective classification testifies to Mary's direct acquaintance with some distinct phenomenal property. Before she experienced red, she had never been in contact with this phenomenal property. But now that she is acquainted with it, she can classify new experiences depending on whether they display it or not.\(^6\)

Once more, however, there is an obvious materialist story to be set against this, which accounts equally well for the fact that people can only introspectively classify into kinds they have previously experienced, and does so without invoking any special phenomenal properties.

Suppose that introspective classification depends on the existence of some kind of brain "template", to use David Lewis's phrase (1983). We don't classify new experiences by seeing whether they have some phenomenal property that we have previously been acquainted with. Instead we simply compare them with the "template" to see whether they correspond. This hypothesis too yields an obvious materialist explanation of why you should only be able to introspectively classify experiences of a kind you have previously had. Again, the brain needs an original to form the mould. In order to fix a neural pattern as a template against which to compare new inputs, we need to have been some original experience to create the pattern.

To make this template hypothesis more concrete, we might suppose that, whenever the relevant classificatory question arises, the template "sends down" neural signals to

\(^6\) Again, there are puzzles about the role of direct acquaintance in the non-materialist story. How does it help to classify new experiences that you were previously acquainted with some phenomenal property? The questions pressed about imaginative recreation in the previous footnote apply equally to introspective classification.
lower levels of perceptual processing. A positive classification would then be triggered by a "match" between these backwards signals and current sensory input. This match could then boost the activation of the template, and this boosted activation could itself serve as the relevant classification of current experience.

Now, the precise accuracy of this picture is clearly hostage to empirical research. Still, if anything even roughly along these lines is right, it will yield a natural explanation of why Mary's experience should enable her to think in ways she couldn't think before, and moreover an explanation which doesn't require any distinct phenomenal properties.

1.6 The Ability Hypothesis

On the inflationist view, when Mary comes to "know what seeing red is like", she acquires a new concept of seeing red, a phenomenal concept, which is distinct from any material concepts she previously possessed. She mightn't be acquainted with any new phenomenal property—some of her old material concepts already referred to the property of seeing red—but she has a new way of thinking about that property.

It might not yet be clear, however, why I want to count the before-after changes in Mary as amounting to her acquisition of a new concept. We have seen how she will have new powers to imaginatively recreate and introspectively classify red experiences. But why view these changes as the acquisition of a concept?

However, note that Mary's new powers will also enable her to think new certain kinds of thought. Now she can imagine red experiences, Mary can think thoughts like "People looking at ripe tomatoes experience that". And her new introspective powers will also allow her to think thoughts like "This is what people experience when they look at ripe tomatoes". That is, she will be able to deploy her imaginative and introspective powers in the construction of articulated thoughts, mental judgements that can be true or false.

When I speak of concepts, I mean components in just such truth-evaluable thoughts. Concepts are elements which make a systematic contribution to the truth conditions of the thoughts they enter into. By this criterion, the "this" and "that" in the thoughts attributed to Mary above would seem to represent something conceptual. For the overall thoughts containing the "this" and "that" certainly seem to be evaluable as true or false.

Some philosophers are happy to accept that Mary acquires new powers of imaginative recreation and introspective classification, yet deny that it is appropriate to view this as a matter of her acquiring any new phenomenal concepts. These are the sophisticated deflationists mentioned earlier. They accept, contra Dennett, that some genuine and lasting before-after differences would be occasioned in Mary by her new experience. In particular, they accept that she will now be able to recreate that kind of experience in imagination, and be able to classify new experiences introspectively as of that kind. Yet they deny that Mary will thereby become able to think any new thoughts. What she has are some new ways of forming thoughts, but the thoughts formed will all be of
the old kind that she could think even before she came out of her house. (Lewis, 1988; Nemirow, 1990.)

Now, there is a sense in which even inflationists like myself want to deny that Mary can think any new thoughts. On our view, Mary doesn't think about any new entities. The conscious property she can now think about as "that experience", once she has seen something red, is just the same property as she could always think about in material terms. So if we type thoughts at the level of reference, so to speak, in terms of what they are about, Mary doesn't have any new thoughts.

Still, on the inflationist view, there is another sense in which Mary does have new thoughts. She may still be thinking about the same old property, but she is thinking about it using a new concept. Typed at the level of sense, rather than reference, she is thinking a new thought. The sophisticated inflationist defenders of the ability hypothesis want to deny even this much. They do not allow that Mary has any new concepts. Even at the level of sense, Mary repertoire of thoughts is not augmented when she comes out of her house and has her first red experience.

It may seem that, in conceding a substantial before-after difference in Mary, the ability hypothesis has no option but to concede that she acquires a new concept. If she comes to know something she didn't know before—she comes to "know what seeing red is like"—then doesn't it immediately follow that she must have some new thought, if only at the level of sense?

Not necessarily, according to the ability hypothesis. For knowledge can include knowledge how, as well as knowledge that. When I find out how to ride a bicycle, I come to know something I didn't know before. I now know how to ride a bicycle. But this needn't involve me having any new thoughts—it needn't involve me knowing that anything I didn't know before. After all, I may already have known everything of a propositional kind about riding bicycles, even before I learned how to ride one. I could have been an absolute expert on the physics, physiology, economics and history of bicycle-riding, and just not have acquired the knack myself. If so, what I would have lacked was not any kind of thoughts about bicycle riding, however typed, but simply the ability to ride a bicycle myself.

So it is with Mary, according to the ability hypothesis. On a sense, she did not "know what seeing red is like" before she came out of her house. But she wasn't incapable of thinking any thoughts, however typed, about red experiences. All she lacked was the ability to recreate that experience in imagination, and the ability to classify it by introspection.

As a version of deflationism, the ability hypothesis is clearly preferable to Dennett's outright denial that Mary in any sense comes to "know something new". It allows that Mary acquires new powers of imagination and introspection. Even so, the ability hypothesis seems to me not to do justice to the change in Mary. If we look more closely at Mary's new abilities, we will see that they are inseparable from her power to think certain new kinds of thoughts.

For a start, note that these new abilities, unlike the ability to ride a bicycle, are at least closely connected with certain kinds of thought. As I put it earlier, Mary's new
exercises of imagination and introspection will be accompanied by thoughts like "People looking at ripe tomatoes experience this". At first sight, this doesn't look akin to Mary moving her arms and legs on a bicycle, but rather to her forming a new thought of a kind she couldn't form before she left her house.

Perhaps this is too quick. Ability theorists might respond that while Mary's new abilities are indeed thought-involving, unlike riding a bicycle, they aren't new thought-involving.

To take introspective classification first, the idea would be that Mary has a new introspective route to beliefs involving the concept of experience of kind Ø. She can now arrive at beliefs with this content directly and introspectively, whereas before she could only ascribe an "experience of kind Ø" to people on different grounds, by knowing about their behaviour, say, or by knowing about physical goings-on inside them. But there isn't any new concept here. The concept of an "experience of kind Ø" will be some old material concept she had before she left the house. She's just acquired a new technique for applying the concept.

And a similar line might be taken with the thoughts involved in Mary's new imaginative abilities. Mary recreates the experience of seeing red in her imagination, and simultaneously thinks something like "Ripe tomatoes cause that". But perhaps the concept expressed here by "that" is simply identical with some old material concept of an experience of kind Ø that Mary always possessed. Her thought involving this concept is now accompanied by an act of imaginative recreation, an act that Mary couldn't have performed before she had her first red experience. But the thought itself involves only concepts she always had available.

Once spelled out, the problem with this line is obvious. Why suppose that the concepts involved in Mary's introspective and imaginative thoughts can be equated with old material concepts she always possessed? When I imagine seeing red, and think "That is caused by ripe tomatoes", I certainly don't seem to be deploying any particular material concept. Nor do I seem to be doing so when I introspectively think "this is caused by ripe tomatoes".

To drive the point home, note that Mary may not yet know which of her old material concepts applies to her new experience. Imagine that she is shown, not a rose, but a red sheet of paper, so she has no way of knowing, in her old material terms, which colour experience this is. She might be able to figure out that it is a colour experience, but there is nothing to tell her whether she is seeing red or green or blue. This shows that Mary cannot be thinking just using her old material concepts.

Suppose that Mary, after being shown the piece of red paper, uses her new imaginative powers to hazard "I'll have that experience again before the day is out". This is clearly a thought in full working order—after all, it will either be true or false—but equally clearly it is not equivalent to any thought Mary can form using her old material concepts, since she has no idea which of those picks out "that experience". Again, suppose that this same Mary later confirms her guess with the introspective classification "Now I'm having this experience again—and it's not yet 3 o'clock". As before, this classification can't possibly be equivalent to any classification using one of
Mary's old material concepts. since again she will not know which of these picks out "this experience".7

There is a sense in which Mary's new powers of imaginative recreation and introspective classification are indeed new abilities—she can certainly do things she could not do before. But they are not mere abilities, if that is taken to rule out her possession of new concepts. At the level of reference, Mary may still be thinking about the same properties she could always think about. But at the level of sense her new imaginative and introjective powers generate a new way for her to think about those properties.

### 2.7 The Contingency of Learning from Experience

In some ways the inflationist account of phenomenal concepts is reminiscent of the traditional empiricist account of ideas. Hume maintained that all ideas are copies of impressions. You can only think with concepts that you have derived from earlier experiences. Inflationists need not of course agree with Human that all concepts have such an experiential source. Inflationism is a thesis specifically about phenomenal concepts of conscious properties, and carries no implication about non-phenomenal concepts of conscious properties or of anything else. But when it comes to phenomenal concepts themselves, inflationists do maintain that you need an initial experience, as with Mary, to acquire a phenomenal concept.

Of course, some obvious qualifications are needed, analogous to those orginally made by Hume about complex and intermediate experiences. It is possible to know what complex experiences are like—that is, to be able to imaginatively recreate them and to introspectively classify them—even if you have never had them before, provided you have had the simple experiences out of which these complex experiences are composed. For example, we are capable of imaginatively recreating and introspectively classifying the complex experience of seeing a unicorn, as long as we've previously experienced the elements of seeing a horse and seeing a creature with a horn. In addition, it is arguable that we are sometimes capable of imaginatively creating or introspectively classifying intermediate experiences we haven't had before, provided we have previously had relevantly related experiences. For example, we might sometimes be able to imagine or classify a colour experience which is spectrally intermediate between other colours we have previously experienced.

I shall take these qualifications as read from now on. Given this simplification, it interesting to ask why you can only "know what an experience is like" once you have had it yourself. The materialist will answer this question quite differently from the anti-materialist.

The anti-materialist view holds that "knowing what it is like" requires some kind of acquaintance with a non-material property. On this view, it will be a matter of necessity that you must undergo an original experience before you can acquire the corresponding phenomenal concept. You can't have the concept unless you are

---

7 For similar arguments against the ability hypothesis, and for conceptual dualism, see Loar, 1990.
acquainted with the property, and you can't be acquainted with the property unless you have experienced it at first hand.

On the materialist account of phenomenal concepts, by contrast, it comes out as a quite contingent matter that you need an original experience to "know what it is like". It does seem that humans beings, subject to the qualifications above, can only think phenomenally about experiences they have had before. But the explanation I have suggested for this phenomenon implies that it could well have been otherwise, with different kinds of creatures.

I argued that Mary could not imagine or introspectively classify experiences of red without an original experience because, as I put it, the brain needs an "original" from which to make a "mould". The relevant patterns of neural activation can only initially be fixed by an exogenously caused experience. Now, while this may indeed be true of us humans, it easy enough to posit creatures who do not work like this, and so are not subject to the same cognitive limitations. These would be creatures who are born with imaginative and introspective abilities, so to speak, and do not need any specific experiences to instil them. The necessary "moulds", and the dispositions to use them, would be "hard-wired", that is, grow independently of any specific experiences. A creature who developed like this would be able to imagine seeing red, and be poised to classify new experiences as of that type, even before they underwent any exogenously caused red experience. Humans are not like this. But there seems nothing impossible about creatures who are.

2.8 Imagination and Introspection

Let me now focus on the relation between imaginative recreation and introspective classification. So far I have taken it that these two powers go hand in hand, that they are simply two sides of the same coin of "knowing what it is like".

Still, on reflection, it does not seem inevitable that the two abilities should accompany each other. There seems nothing incoherent in the idea of creatures in whom they come apart, that is, who can imaginatively recreate experiences, but not introspectively classify them, or who can introspectively classify experiences, yet never imaginatively recreate them. To make these possibilities concrete, simply posit having some pattern of neural activation that can be used for introspective classification, yet the lack of anything similar that can be switched on in imagination, or vice versa.

Indeed we need not go to imaginary creatures to find such dissociations. While it does seem plausible that humans can introspectively classify everything that they can imagine, the link the other way seems far less tight. For example, we seem much better at introspectively classifying smells than at imaginatively recreating them. An actual olfactory experience can create an intense feeling of recognition, yet we may be quite unable imaginatively to recreate that smell later. And to some extent the same point applies across the experiential board. We are rather better at recognizing experiences than recreating them.

Still, the two kinds of power do tend to accompany each other in humans, which is why it is natural to lump them together under the heading "knowing what it is like".
Why should this be so, given that it is principle possible, and to some extent actual, for the two powers to become dissociated? Smells and similar examples show that the correlation between imagination and introspection in humans is by no means perfect. But there remains a question as to why there should be any correlation at all.

An obvious answer is suggested by the models of imaginative recreation and introspective recognition suggested earlier. Perhaps the same mechanism underlies both powers. That is, perhaps the patterns of neural activity which are "switched on" in imaginative recreation are just the same patterns which provide the "template" for incoming neural signals in introspective classification. So, once some such neural pattern has been fixed by an original experience, it will become available for deployment both in imagination and introspection.

This model also suggests a natural explanation of why we aren't always as good at imagining things as we are at introspective classifying them. We can suppose that, in the standard cases of introspective classification, the "template" neural pattern will be activated automatically by some initial match with incoming neural signals. In line with the model suggested earlier, the template might then "send down" further signals, to check whether this initial indication of a match can be "filled out".

By contrast, neural activation will not be so directly triggered in the paradigm cases of imaginative recreation. Rather, acts of imaginative recreation will result from deliberate choices, or associative connections with other experiences, or from other such sources. If this is right, then we can expect some neural patterns, in some individuals, to be regularly triggered by incoming signals, thus yielding introspective classification, yet to be unavailable to imaginative recreation. This would happen if there were no links allowing deliberate choices, associated experiences, or anything similar, to excite the relevant neural pattern. Perhaps this is why most of us are no good at imagining smells. We can identify smells all right when we have them, but we have no other way of turning on the relevant pattern of neural activation.

2.8 Further Issues

Let me now briefly draw attention to some further questions involving phenomenal concepts. This will enable me to flag some issues which I shall return to later.

2.8.1 Are Phenomenal Concepts Introspective or Imaginative?

If introspective classification and imaginative recreation are separate powers, which can come apart in principle, and to some extent in practice, then ought we to speak of single phenomenal concepts of types of experience? For example, I can think about the experience of seeing something red in virtue of introspectively classifying it, and I can think about it by imaginatively recreating it. So shouldn't we speak here of two distinct phenomenal concepts, an introspective concept and an imaginative concept?

I don't think anything much hangs on this question. It is more a matter of how we describe our data, rather than anything substantial. Accordingly, I shall continue to talk about phenomenal concepts as such, but will take care to distinguish between imaginative and introspective deployments of these concepts when it matters.
2.8.2  Perceptual Concepts and Phenomenal Concepts

Later on, in chapter 4, I shall compare phenomenal concepts with perceptual concepts. These are concepts, not of experiences of seeing red, or seeing an elephant, or feeling a circle, but of redness, or elephants, or circles, considered as denizens of the non-mental world "out there". I shall argue that there is an intimate relation between phenomenal concepts and perceptual concepts.

Because of this relation, there will be a number of analogies between phenomenal and perceptual concepts. At this stage let me just observe that perceptual concepts, like phenomenal concepts, can be variously deployed both in perceptual classification and perceptual recall. So in this case too there will be some reason to speak of two concepts, a classificatory concept and a recollective one. But again this will not be a substantial issue, but simply a matter of how we describe our data.

2.8.3  Indexicality

My uses of "this" and "that" to express Mary's new phenomenal thoughts might have suggested that her new mental powers simply derive from her new opportunity to refer indexically to her own red experiences—when before she couldn't, for lack any such experience to indicate. On this account, she wouldn't have any new concepts, just a new context in which she can now her use old indexical constructions.

However, this simple story doesn't stand up, as will be shown at length in chapter 4. Phenomenal thoughts may involve a kind of sui generis "pointing", and this may explain why they are often expressed by such phrases as "this" and "that". But, even so, it will turn out that this "pointing" is no ordinary indexical device.

2.8.4  Theories of Reference

If phenomenal concepts don't refer in virtue of some ordinary indexical content, how do they refer? How is it possible for us to refer to conscious experiences by exercises of imagination and introspection? This is a substantial question, which will be crucial for much of what follows. The next chapter, which addresses Saul Kripke's well-known modal argument against materialism, will place constraints on possible answers to this question. (In particular, it will show that phenomenal concepts can't refer by description.) But it will leave it open how they do refer, and we shall return to this question at length in chapter 4.
Chapter 3
10
Kripke's Modal Argument and Direct Phenomenal Reference

3.1 Introduction

The conceptual dualism introduced in the last chapter distinguishes sharply between phenomenal and material concepts. The referents of phenomenal concepts may be normal material properties, of a kind that can also be referred to by material concepts. But the phenomenal concepts themselves are quite distinctive, involving special powers of imagination and introspection.

Property identity claims involving phenomenal and material concepts are themselves quite distinctive. Cases where two concepts refer to one entity are familiar enough, and for the most part philosophically well-understood. By contrast, there is something very odd about the phenomenal-material identity claims advocated by materialists. When materialists urge that seeing red (and here you must imagine the redness) is identical to some material brain property, it strikes many people that this must be wrong. "How can technicolour phenomenology arise from soggy grey matter?" asks Colin McGinn (1991), and his question will no doubt strike a chord with many readers.

Materialists will respond that phenomenal-material identity claims may be unusual, but this is no reason why they cannot be true. Indeed, they can point out, the arguments in chapter 1 give us every reason to identify phenomenal properties with material ones. True, when we looked at those arguments in chapter 1, we paid no attention to the special nature of phenomenal concepts. But that didn't matter, materialists will insist, since those arguments didn't depend on any special assumptions about the nature of phenomenal concepts. They simply appealed to a number of compelling empirical claims, which we have as yet seen no reason to deny.

So materialists can with some reason maintain that the onus of argument lies with their opponents. They can agree that phenomenal-material identity claims are not like other identity claims. Moreover, they can recognize that many people are initially disinclined to accept them. But oddity and unpopularity are in themselves no reason to reject theories with substantial empirical backing.

In this chapter I shall look at one anti-materialist argument which goes beyond accusations of oddity and unpopularity. This is Saul Kripke's modal argument (1971, 1972, 1980). Described crudely, it tries to show why the oddity of phenomenal-material identity claims is indeed a reason to reject them.

I shall argue that Kripke's modal argument does not work. But it is worth examining it in some detail. Even if it does not establish its intended conclusion, it will add substantially to our understanding of the workings of phenomenal concepts.

3.2 Epistemology versus Metaphysics
The initial target of Kripke's argument was early post-war materialism, as defended by figures like U.T. Place (1956) and J.J.C. Smart (1959). These early materialists were fond of saying that the identification of mental states with brain processes is a contingent identity. By this they meant to convey that the identification rested on empirical evidence, and could not be established by conceptual analysis alone.

By way of analogy, they invoked scientific identifications like that of temperature with mean kinetic energy, or lightning with atmospheric electrical discharge, or water with \( \text{H}_2\text{O} \). Obviously, there is no question of establishing these things by conceptual analysis alone. A priori reflection on concepts is not going to tell us that temperature is mean kinetic energy. Nevertheless, scientific investigation has shown us that this is indeed so, and similarly with the other identities. True, the scientific results could have pointed to different conclusions. But they didn't. Similarly with mind and brain, said the early materialists. There is no a priori way of showing they must be identical. But, as a matter of contingent scientific fact, it turns out that they are.

Kripke objected that this doctrine of contingent mind-brain identity is confused. The early materialists were confusing the epistemological question of whether mind-brain identities can be established by a priori means alone, or only a posteriori, with the modal or metaphysical issue of whether the claims thus established are necessary, or only contingent. There is of course nothing wrong with insisting that the relation between mind and brain is an empirical matter, to be assessed in the light of the empirical evidence, and not on a priori grounds. But this in itself, insisted Kripke, leaves the modal status of the materialists' claims quite open. There is no legitimate inference from a claim being a posteriori to its being contingent.

Indeed, continued Kripke, once we separate the metaphysics from the epistemology, we can see that the materialists' claim of mind-brain identity would have to be necessary, if it were true at all. This is because all identities are necessary. A thing is what it is, and cannot be something else.

It may be a matter of empirical discovery to find out that Cicero is identical with Tully, say, or the Evening Star with the Morning Star, or temperature with mean kinetic energy. But the truths so discovered are necessary truths. To suppose otherwise is to suppose that Cicero might not have been Tully (or the Evening Star might not have been the Morning Star, or temperature might not have been kinetic energy). However, these things make no sense. How could Cicero possibly not have been Tully? There is no possible world where Cicero exists, but not Tully. Since they are the same person, Tully will be there if Cicero is. Similarly, you can't have a world with the Evening Star but not the Morning Star, or with temperatures but no mean kinetic energies.

So identities are necessary, if true. In particular, mind-brain identities would have to be necessary, if they were true.8

---

8 The claim that all true identities are necessary is often qualified by a proviso about rigid designators. I aim to avoid these complications by making the assumption that all genuine identity claims have rigid designators on both sides. Claims like "John is the tallest man in England" will be assumed not to be identities, but to have the quantificational form (Ex)(x is the tallest man in England and x=John). I realise that some philosophers hold that there can be genuine identities involving
3.2 The Appearance of Contingency

So far this mightn't look like much of an objection to materialism. For why can't materialists simply accept Kripke's distinction between epistemology and metaphysics, and agree that mind-brain identities are necessary, while based on empirical evidence? After all, the important point, for the early materialists, was simply that mind-brain identities are an a posteriori, empirical matter. Kripke shows that it is wrong to muddle this up with these identities being contingent. So the obvious solution is for materialists to disentangle their metaphysics from their epistemology, and simply agree that their a posteriori identities are necessary.

The trouble now, however, is that these identities don't seem necessary at all. The necessity of Tully = Cicero shows itself in the fact that a world containing Tully but no Cicero makes no sense. But there doesn't seem anything similarly incoherent about a world with pains but no brains, or brains but no pains. It seems possible for pains and brains to come apart, in a way that Cicero and Tully simply can't.

Let us suppose, for the sake of the argument, that the materialist wants to identify pains with pyramidal cell activity. Then, by analogy with the Cicero-Tully case, it ought to follow that there are no possible worlds where there is pyramidal cell activity but no pains, or pains but no pyramidal cell activity. But these things seem manifestly possible. In the actual world, these two states may never come apart. But there doesn't seem anything metaphysically incoherent about creatures who are physically just like us, down to their pyramidal cell activity, but who have no feelings of pain. Even less does there seem anything incoherent about a possible world where there are beings who feel pains, but have no active pyramidal cells.

You might feel that these intuitions of possibility simply reflect the impausibility of identifying pains with pyramidal cell activity, rather than with some more abstract or higher-order material property. But this doesn't wash. It doesn't matter which material property you choose as the candidate for identity with pain (or for identity with whichever other conscious property you may be interested in). It will still seem possible for the conscious feeling and the material property to come apart. To see this, we need only consider the possibility of zombies and ghosts.

Zombies are beings who share all our material properties, yet have no consciousness whatsoever. Zombies seem metaphysically coherent, even if never actual. Just imagine a being who is a molecule-for-molecule duplicate of yourself, but who feels nothing at all, who is a mere automaton so far as conscious experience goes. Of course, we don't expect ever to meet such a being. Actual people don't work like that. But, still, there seems nothing incoherent about such an insensate doppelganger, who has all your material attributes, yet lacks the conscious ones.

Ghosts are the converse possibility—beings who share none of our material properties, yet have just the same conscious states as we do. Again, ghosts seem metaphysically coherent, even if never actual. Just imagine a being who shares your conscious life, yet flaccid designators. But nothing in my arguments depends on this issue, as far as I can see, and it will be far more convenient to assume all identities are necessary.
has no material properties at all, of the kind which underpin conscious life in this world. Such a being would share all your conscious properties, yet have none of your material properties.\footnote{For a discussion of some philosophical asymmetries between zombies and ghosts, see Sturgeon, forthcoming.}

If zombies and ghosts are possible, then phenomenal properties cannot be identical with any material ones. Take a generic conscious property \(C\). The possibility of zombies and ghosts implies that it is possible for \(C\) to come apart from \(M\), for any material property you may wish to identify \(C\) with. But if this is possible, then it follows that \(C\) cannot be identical with any material \(M\). For the dissociation would not be possible if \(C\) were really identical with \(M\), any more than it is possible for Cicero to come apart from Tully.

Kripke's argument is thus that the possibility of conscious properties coming apart from material properties shows that they cannot be identical with material properties. Kripke can of course allow that certain conscious properties are always found hand-in-hand with certain material properties in the actual world. But, from Kripke's point of view, this will only mean that those properties are correlated, not that they are identical. The properties can't be identical, for then there would be no sense to the idea that they might come apart—which there clearly is, insists Kripke.

(Of course, if Kripke is right here, and conscious properties really are distinct from material ones, then the arguments of chapter 1 show that either (a) physics is causally incomplete, or (b) epiphenomenalism is true, or (c) the effects of conscious states are invariably overdetermined. Some of the philosophers moved by Kripke's arguments have shown themselves prepared to embrace one or the other of these options.)

### 3.3 Explaining the Appearance of Contingency

Since materialists are committed to mind-brain identities, and identities are necessary, they need to deny that conscious states can possibly come apart from material ones. It is difficult, of course, to deny that these things \textit{seem} possible. There doesn't \textit{seem} anything metaphysically incoherent about the possibility of zombies and ghosts. But materialists must deny that such things really are possible. So they need to say that zombies and ghosts are a kind of modal illusion. Even though it might seem to us that conscious and material states can come apart, such dissociations are not really possible.

However, materialists now face another challenge. Why should zombies and ghosts \textit{seem} possible, if they are not? On the face of it, the mind-brain relation seems quite different from other identities, like Cicero=Tully, precisely in appearing contingent where they do not. Materialists say that this appearance is illusory. But then they surely owe some explanation of this illusion. Since they agree that the mind-brain at least \textit{seems} to be contingent, they need to come up with some explanation for the appearance of contingency.

Still, it might seem as if an explanation is ready to hand. Go back to the scientific identities that the early materialists originally held up as their model of “contingent
identities", like temperature is mean kinetic energy, or lightning is electrical discharge. Now, we can all agree that, give that these are indeed identities, they can't really be contingent, however much they are a posteriori results of scientific investigation. But, still, don't they at least appear contingent?

On the face of things, it makes perfectly good metaphysical sense that there should be worlds in which there are temperatures, but no mean kinetic energies, or mean kinetic energies, but no temperatures. After all, aren't these just of worlds which science might have shown to be actual, even though it didn't? For example, what about a world in which sensations of heat turn out to be caused, not by mean kinetic energy, but by the flow of some distinct caloric fluid? Or what about a world in which there are mean kinetic energies all right, but our perceptual apparatus works rather differently, so as to stop us registering mean kinetic energies as sensations of heat? These certain look like metaphysically coherent worlds in which temperature and mean kinetic energy come apart.

Of course, given that temperature actually is mean kinetic energy, then the arguments of this chapter so far show that these aren't really worlds in which temperature and mean kinetic energy come apart. If these two quantities are the same quantity, it can't come apart. Rather, the relevant worlds are ones in which something other than temperature (that is, other than mean kinetic energy) causes heat sensations, or, alternatively, in which temperature (that is, mean kinetic energy) doesn't cause heat sensations.

Still, this last point is no real hindrance to the use materialists want to make of the analogy with temperature and mean kinetic energy. For their aim is to explain the appearance of mind-brain contingency, not its actuality. And temperature and mean kinetic energy would still seem to provide a perfectly good model for this. For it still appears as if they could come apart, even if this isn't really possible. After all, it is surely quite natural to describe the relevant worlds in the terms I first used, namely, as worlds in which there are "temperatures, but no mean kinetic energies", or "mean kinetic energies, but no temperatures". This may be loose talk, but it is very natural talk, and this in itself surely suffices to explain the common, if confused, impression that temperatures might not have been mean kinetic energies, or vice versa.

So the idea would be to offer a similar explanation for the apparent contingency of mind-brain identities. Materialists can argue that these identities strike us as contingent because we are aware that science might have uncovered rather different facts from the ones it actually did uncover. This is why we confusedly think that "temperature might not have been mean kinetic energy". Similarly, so the materialist suggestion would go, with the thought that "pains might not have been pyramidal cell activity (or any other material state)".

3.4 Referring Via Contingent Properties

The Kripkean argument isn't finished yet. We need to look more closely at the analogy with temperature and mean kinetic energy. It turns out that it isn't as helpful to the mind-brain materialist as it seems.
Let us consider which world we are actually describing when we posit a world in which, say, "mean kinetic energy is not temperature". This is the world in which we have different perceptual mechanisms, and so mean kinetic energies fail to cause sensations of relative heat. Now, the reason we naturally describe this as a world in which mean kinetic energy is not "temperature" is that we initially think of temperature as that quantity, whatever it is, that causes sensations of relative heat. So, when we specify a world in which mean kinetic energy fails to satisfy that description, it is natural to describe it as one in which mean kinetic energy is not "temperature". This is a misdescription, given that "temperature" actually is mean kinetic energy—the relevant world is more properly described as one in which mean kinetic energies do not "cause heat sensations". But it is still a very natural misdescription.

Now, we can expect to find this set-up with most scientifically established identities. The pre-theoretical terms involved will have their references fixed by description. Thus, everyday terms, like "water", or "temperature", or "lightning", will pick out that quantity or property which satisfies some everyday description. Prior to the scientific investigation, we won't yet know which property this is, that is, we won't yet know that the relevant term names H2O, or mean kinetic energy, or electrical discharge. These things are for science to discover. So initially the reference of the everyday terms will be fixed via some pre-theoretical description, like "odourless, colourless and tasteless liquid", "causing heat sensations", or "flashing through the sky before thunder". These descriptions will be associated a priori with our initial terms.

However, the properties these descriptions invoke (causing heat sensations, and so on) will only be contingently possessed by the referents. Temperature (that is, mean kinetic energy) has the property of causing heat sensations contingently. This shows itself in the fact that there are genuinely possible worlds in which mean kinetic energies do not cause heat sensations, because of alterations in our sense organs. These are the worlds which we are tempted to describe inaccurately as having "mean kinetic energy but not temperature", though strictly they are only ones where "mean kinetic energy fails to cause heat sensations".

The important point in all this is that we only get an appearance of contingency with scientifically established identities because the everyday terms involved in such identities have their references fixed by contingent properties. Because these properties are contingent, there are genuinely possible worlds in which the scientific referents lack these properties, such as, for example, worlds in which mean kinetic energies do not cause heat sensations. Because these properties fix reference, it is natural to describe these as worlds in which the scientific referents come apart from the everyday terms, as worlds in which mean kinetic energy is "not temperature", even though, strictly speaking, this is an inaccurate description.

This suggests that, if materialists are to run the same story with mind-brain identities, they will have to hold that pre-theoretical terms for conscious states, like "pain", pick

---

10 When I talk about "reference by description", as I frequently shall from now on, I should be understood in a generous sense. I make no assumption that the relevant term's reference is, or was, fixed by explicit stipulation, nor even that the description can be articulated in language. I shall mean only that the term is a priori equivalent to something of the form "the (actual) x which is Ø", where "Ø" represents some independently referring concept.
out their referents via contingent properties. This is where Kripke's argument really bites. For, when we try to run this model, it turns out not to work.

On this model, "pain" will have to fix its reference to some material property M via some contingent feature of that property. The idea would need to be something like this: the material property M, the real referent of "pain", is picked out as that property, whichever it is, that contingently generates painful reactions in humans. Given this, there would then seem to be space for a world in which pain (that is, M) does not generate those painful reactions—for example, a zombie world in which there are beings who share our Ms but not our painful feelings.

And then the explanation for the apparent contingency of mind-brain identitires would need to run as follows. "It may be natural to describe the zombie world as one in which there are 'Ms but not pains'. Indeed this explains our impression that the relation between pain and M is contingent. But this relation is not really contingent, for we are misdescribing the relevant world: the zombie world is not one which lacks pains—it just lacks the further contingent property by which we pick out pains in this world, namely the generation of painful reactions."

Unfortunately, something seems to have gone wrong here. For surely the zombie world lacks pains, not just "painful reactions". For what are pains, except the "painful reactions generated in humans"? The "generation of painful reactions" can't plausibly be viewed as some contingent property which helps us to pick out pains in the actual world. Surely it is the essence of pain itself.

Stubborn materialists may feel inclined to dig in their heels here, and insist that zombies do have pains. That is, they could insist that pain itself is different from "painful reactions". Pain is identical with some material state, and so is present in zombies. It just fails, in the zombie world, to generate those subjective "painful reactions" with which pain is contingently associated in the actual world, and which we happen to use to fix the reference of our word "pain".

But this ploy not only requires a quite implausible account of the working of the concept "pain"—it doesn't help anyway. For anti-materialists can now simply switch their attack to the "painful reactions" themselves. It is agreed on all sides, and in particular by the materialists, that the zombies lack these, even though they are experienced by humans who are in pain in the actual world. Anti-materialists can simply point out that this in itself refutes materialism. If all conscious properties are material, then how come zombies, who are stipulated to share all the material properties of humans, so much as lack "painful reactions"? Materialists need to identify "painful reactions" themselves with some material property, and so can't coherently suppose that zombies lack them.

So, whichever way they turn it, materialists seem unable to offer a satisfactory account of the apparent contingency of mind-brain relations. It looks as if they have no option but to admit that these relations really are contingent: even if properties like pain are perfectly correlated with certain material properties in this world, this correlation could fail to obtain in other possible worlds. But then it would follow that conscious
properties aren't [*identical*] with material properties. For identities, unlike correlations, cannot come apart in other possible worlds.\(^{11}\)

### 3.5 A Different Explanation

Materialism may be temporarily down, but it is by no means out. There is another way to answer Kripke's challenge.

There is nothing wrong with most of Kripke's argument. Identities are indeed necessary. So materialists must deny that it is possible for conscious properties to come apart from the material properties they are identical with. At the same time, it certainly *seems* possible that these properties should come apart. So materialists owe an explanation of this appearance of contingency. Yet it won't do to say that phenomenal concepts like "pain" pick out their referents via contingent descriptions, a la "temperature". This claim itself turns out to be inconsistent with materialism.

The loophole is that the contingent description story isn't the only way to account for the appearance of contingency. The materialist can agree with all the Kripkean points listed in the last paragraph, yet offer a different explanation for the appearance of contingency, one which doesn't have phenomenal concepts referring by contingent description.

Before explaining how this might work, it will be worth emphasising exactly why the contingent description story is a poisoned chalice for materialism. At its most graphic, the challenge facing materialism is to explain why zombies seem not to satisfy the concept "pain", even though by materialist lights they should. Anti-materialists have an obvious answer—"pain" refers to some non-material property which zombies lack (and so zombies don't just *seem* not to satisfy "pain"—they don't). Obviously, materialists need some other answer. However, as soon as they accept Kripke's invitation to assimilate the concept "pain" to concepts that refer by description, they are in trouble. For then they need to explain why zombies do not satisfy whichever concepts are involved in the relevant description. And this takes them back to where they started, with no obvious alternative to the anti-materialist claim that zombies simply lack the non-material properties present in genuinely sentient beings.

---

\(^{11}\) There might seem to be another way for materialists to appeal to the idea of reference by contingent description. Why not apply this idea to the *material* concepts involved in the mind-brain identities, rather than to the phenomenal terms? After all, it is quite natural to hold that such concepts refer to physical kinds as entities which play specified theoretical roles. So can't the materialist portray a zombie world as one where those theoretical roles are filled by something other than the stuff which fills them in the actual world? *Pain* could then be identified with something involving that actual stuff, and zombies wouldn't feel pain because they lacked that actual stuff, while satisfying the contingent descriptions which pick it out in this world.

Well, I am happy enough to accept that such creatures are possible—that is, beings that lack pains because they lack the actual stuff picked out by scientific descriptions in this world. But I don't think that it helps us to explain the apparent possibility of zombies. For by my lights these creatures aren't a proper zombie. This is because I take the stuff picked out by scientific descriptions in the actual world to be physical stuff (cf, footnote 5 in chapter 1). So proper zombies will share this stuff, as well as satisfying the descriptions which pick it out—and still lack pains. And the materialist will still owe an explanation of why even such proper zombies seem possible.
Materialists should simply refuse Kripke's invitation. They should say that phenomenal concepts refer directly, and not by description. Phenomenal concepts don't pick out their referents by invoking certain further features of those referents, but in their own right, so to speak.

This claim of course raises questions about how phenomenal concepts do this. How do phenomenal concepts pick out their referents, if not by description? But let us not pause to answer this question at this point. It involves a number of issues, which I shall discuss at length in the next chapter. For now it is enough to note that materialists must assume phenomenal concepts somehow refer directly, if they are to avoid Kripke's trap.

Given this, materialists can aim to construct a different explanation for the appearance of mind-brain contingency. To start with, they can point out that we have these two quite different ways of referring to conscious properties. We can refer to them both with phenomenal concepts and with material concepts. And this in itself, materialists will maintain, can generate the impression that phenomenal properties might be different from material properties.

After all, materialists can point out, the distinctness of phenomenal and material properties certainly makes it conceivable that zombies and ghosts should exist. Since there are no a priori connections between phenomenal and material properties, there is no conceptual contradiction in positing beings with all relevant material properties but no conscious ones, or vice versa. Materialists must of course deny that the conceivability of these things shows they are really possible. But they can still maintain that it shows why they seem possible. Zombies and ghosts seem possible, materialists can thus say, simply because they can be imagined without violating any a priori, conceptual constraints, and not for any other reason.

3.6 Thinking Impossible Things

This might seem a bit quick, and indeed it is. Let me go a bit more slowly. There a number of further issues raised by this materialist response to Kripke.

For a start, some may want to question whether the response really makes sense. Suppose materialists are asked to explain what people are thinking, when they entertain the possibility of zombies and ghosts? What possible worlds provide a content for these thoughts?

In general, when we think something, the content of our thought can be equated with some set of possible worlds, namely, those possible worlds whose actuality would make the thought true. On this model, there is no problem about false thoughts—they are simply those whose content consists of possibilities which do not include the actual world. But impossible thoughts do seem problematic, since there are no possible worlds to give them content.

---

12 When I talk about "direct reference", as I frequently shall from now on, I will simply mean reference that is not by description.
We have already seen how this challenge might be answered in connection with one kind of apparently impossible thought. Consider thoughts like "temperature might not have been mean kinetic energy". Now, this description represents an impossibility. Still, we can regard this description as a misdescription of what is really being thought. That is, we can redescribe the thought as really answering to a genuinely possible world, namely, a world in which it is not mean kinetic energy that causes heat sensations. Similarly with "water might not have been H2O", or "lightning might not have been electrical discharge". In each case, we can "reconstrue" these strictly "impossible thoughts" as laying claim to genuine possibilities, namely, possibilities in which the relevant properties fail to satisfy the descriptions which everyday terminology uses to pick them out.

But note how this story requires that the relevant properties are picked out by descriptions to start with. Without these descriptions, there wouldn't be any real possibilities to breathe content into the "impossible thoughts". So this now gives us a new version of the Kripkean challenge. If phenomenal concepts pick out their material referents directly and without description, as the materialist now has it, then how can we so much as think that consciousness might come apart from material properties? For, on the current materialist story, we would not only be thinking a strict impossibility, but there would be no descriptions around to convert this impossibility into a genuinely thinkable possibility.

In response to this challenge, materialists should simply deny that there is any difficulty about thinking impossibilities. Sometimes our thoughts answer to no genuine possibility. This doesn't stop them being thinkable, even when there is no other possibility around to give them an alternative content. (Alice laughed. "There's no use trying," she said: "One can't believe impossible things." "I dare say you haven't had much practice," said the Queen. "When I was your age, I always did it for half-an-hour a day. Why, sometimes I've believed as many as six impossible things before breakfast.")

By way of analogy, consider a case involving proper names. Suppose Jane is familiar with the names "Tully" and "Cicero", but doesn't know that they name the same person. Suppose moreover, that Jane has no specific beliefs involving these terms. She has picked up the names from other people, and to this extent is competent to use them, but beyond that has no special knowledge of Cicero or Tully. Now, Jane may well entertain the thought that Cicero is not Tully. Indeed she is downright likely to believe that Cicero is not Tully. But note that she will be thinking an impossible thought here. There is no possible world corresponding to her thought, no world in which Cicero is not Tully.

Even so, Jane can surely think this thought. And she can do so even though she doesn't associate any descriptions with "Cicero" or "Tully", and so can't be thinking about some other possible world, some world in which Cicero/Tully doesn't satisfy those descriptions. I take the example of Jane to show that there can be thoughts with impossible contents. Indeed, it gives us a simple recipe for constructing them. Take two names for the same thing, join them in a thought where they flank a term for non-identity—and there you are.
I say the same thing about denials of mind-brain identity. Here we have names for properties, rather than people. But the point is just the same. Two terms—a phenomenal term and a material term—can name the same phenomenal/material property. There is thus no real possibility of non-identity. But this doesn't stop us forming contentful thoughts about non-identity, such as that pain is different from any material property you care to choose. Nothing further is needed to explain the existence of such thoughts. Just take phenomenal concepts, material concepts, and a term for non-identity—and there you are.

Indeed, having come this far, we can see that we may as well have said the same thing about imagining such impossibilities as that temperature is not mean kinetic energy. There is no real need to tell the complicated Kripkean story about our really imagining something else, namely, a world in which mean kinetic energy/temperature lacks the properties which fix reference to that quantity in this world. Why not simply say that "temperature" and "mean kinetic energy" are different concepts, which they clearly are for most people, and use this fact alone to explain how people can think the impossible thought that temperature is not mean kinetic energy without any conceptual inconsistency. The point, once more, is that there is nothing difficult about thinking an impossible thought, once you have two terms for one thing.

Of course, there remains a genuine disanalogy between cases like temperature-mean kinetic energy and the mind-brain cases. Since "temperature" arguably refers via a description, there is indeed is a genuine further possibility—that mean kinetic energy/temperature not satisfy that description—even if we don't need this possibility to provide a content for thoughts that temperature is not mean kinetic energy. By contrast, there is no genuine possibility corresponding to the thought that zombies might have no feelings. Since phenomenal concepts don't refer by description, there is simply no possibility at all in the thought that a being may share your physical properties yet lack your conscious ones.

3.7 Conceivability and Possibility

Some philosophers hold that, contrary to the claims just made, conceivability is a good guide to possibility. They maintain that, to every conceivable non-identity (N  M, say), there corresponds a genuine possibility. In cases where N is M, this can't of course be the possibility that N is not itself. Rather, in such cases, it must be that N (or M) refers by association with contingent descriptions, which then generates the possibility that the entity referred to might not satisfy those descriptions.

Putting all this together, it follows that, whenever N M is conceivable, either (a) N really isn't identical with M, or (b) one of the terms involved refers by description. This makes it clear why materialists must deny the premise. Conceivability cannot be a good guide to possibility. For, if it were, then the manifest conceivability of zombies would imply either (a) that phenomenal properties aren't material properties, or (b) that phenomenal concepts refer by description. The former alternative refutes materialism straight off, while the latter only delays this refutation momentarily, since the
possibility of a zombie not satisfying the relevant descriptions itself refutes materialism.13

My response that conceivability is not a good guide to possibility. I take the Cicero-Tully example from the last section to provide strong support for this view. It is conceivable, for Jane, that Cicero Tully, even though (a) Cicero is Tully and (b) she associates neither Cicero nor Tully with any descriptions.

Someone who wants to uphold conceivability as a guide to possibility will need to argue here that Jane must have some further ideas about Cicero and Tully, if she is to have genuine concepts of them. That is, she must associate certain descriptions a priori with "Cicero" and "Tully", if she is really capable of thinking with these terms. This will then restore the link between conceivability and possibility, since it will give us the possibility that Cicero/Tully not satisfy those descriptions.

But why suppose that any such associations are necessary for Jane to be competent with these terms? The theory of names is a large subject, and this is not the place to start pursuing it. But one clear lesson of the last thirty years of work in this area is surely that Jane's conceptual competence with "Cicero" and "Tully" need owe nothing to any specific ideas she associates with these terms. Rather, it will be enough if she has picked up the names "Cicero" and "Tully" from competent speakers, and intends to use them as they do. And this clearly doesn't require that she associate any further descriptions with these names.

More generally, the contention that conceivability is a guide to possibility places implausibly strong constraints on the theory of reference. It requires that, whenever two directly referring terms refer to the same thing, it must be a priori knowable that they do so. For, on the conceivability -> possibility assumption, if it is so much as conceivable that some directly referring "N" and "M" do not co-refer, then it must be true that NM, for without any associated descriptions there is no other possibility around to explain the conceivability. On the conceivability -> possibility view, then, we can be confident that two entities really are distinct, as soon as they seem distinct when thought about using directly referring concepts.

I see no reason to accept this thesis. It seems to me to hinge on some atavistic view of reference. For it to be true, the basic referential relations, direct referential relations, would have to involve some kind of unmediated mental grasp of the entities referred to, a grasp which left no room for mistakes about identity.14 Far from accepting this, I take the basic referential relations to depend on all kinds of facts external to thinkers' head, facts which create plenty of room for a thinker to be wrong about whether two terms refer directly to the same thing.

Let me conclude this section with a historical observation. There is something ironic in the fact that the works in which Kripke first elaborated the anti-materialist modal argument are also the works in which he first defended the causal view of proper names. For the modal argument is only compelling, as we have seen, as long as there are no impossible thoughts involving only directly referring concepts. Yet, if Kripke's

---

13 For completeness, we should also consider the option that material concepts might refer by description. But we saw in the footnote before last that this goes nowhere.

14 For an extended critique of such ideas of direct reference, see Millikan, 1993.
causal view of names is right, proper names provide the most obvious counter-example to this thesis. I don't know what to make of this curiosity. Perhaps the moral is that it takes time for things to become clear in philosophy, even to the most penetrating minds.

3.8 The Intuition of Distinctness

Let me finish this chapter by drawing attention to one remaining puzzle. So far I have offered an explanation of how zombies and ghosts can seem possible, even though they are not. But I do not think that this explains all the anti-physicalist intuitions to which Kripke's argument draws attention.

As a preliminary to showing this, note that there is a significant disanalogy between the Cicero-Tully case and the mindbrain case. When I introduced Jane as an example of someone who could think an impossible thought, I took care to make her ignorant of Cicero's identity with Tully. She was capable of thinking the impossible non-identity precisely because she had no reason to think Cicero and Tully the same person. Indeed, as I pointed out right at the beginning of this chapter, for someone who does accept that Cicero is Tully, there will cease to be any appearance of possible distinctness. This person will no longer be able to make any good sense of the possibility that Cicero might exist, but not Tully. Since they are the same person, Tully will be there if Cicero is.

This point does not undermine the use I have made of the Cicero-Tully example. Even if the non-identity only appears possible while Jane remain ignorant, the ignorant Jane still gives us an example of someone who can think an impossible thought, a thought to which no genuine possibility corresponds.

Still, the ignorance-dependence of her impression of possibility does mark a contrast with the mind-brain case. For, in the mind-brain case, the impression that mind and brain might come apart is likely to persist, even among those who are persuaded that they must be identical. Take my own case. I would say I am persuaded, by the arguments you are reading in this book, that mind and brain must be identical. Yet zombies and ghosts still strike me as being intuitively possible. I don't seem to have any trouble grasping zombie or ghost scenarios. Isn't this just the idea of phenomenal states without material states, or vice versa?

Given the analysis in this chapter so far, this disanalogy should appear puzzling. Now that I know that Cicero=Tully, can no longer make any good sense of the suggestion that Cicero might not have been Tully. What am I supposed to imagine? That he might not have been himself? But if I can't make sense of this possibility, then I ought not to be able to make sense of zombie and ghost possibilities either. If I accept that pain is identical to some material state, as I do, then oughtn't I to find zombies and ghosts as incoherent as Cicero without Tully? After all, what am I supposed to be imagining? That pain might not have been itself? My analysis so far may have explained why this "impossible thought" will occur to people who do not accept mind-brain identity. But it leaves us with a puzzle about why it should persist in people, like myself, who do.
It don't think that this shows that there is anything wrong with my analysis so far. Rather, it shows that there is some further feature of the mind-brain case that has not yet been taken into account. Here is what I think is going on. There is another source of anti-materialist intuition, which is quite independent of the issues considered in this chapter. This continues to operate even in those, like myself, who are otherwise persuaded that there are good arguments for materialism, and stops us fully believing the materialist conclusion. This is why the mind-brain case is different from the Cicero-Tully case. There is something which stops us ever fully accepting that the mind is identical to the brain, and this is why we continue to think that they might come apart, as in zombies and ghosts.

This extra intuition—which I shall call "the intuition of distinctness"—will be the subject of chapter 6. There I shall aim to explain why this intuition should be so compelling, even though it is false. But at this stage it will be worth clarifying how I think of its relation to the modal considerations discussed in this chapter. I take the intuition of distinctness to be quite independent of any modal matters. It simply a thought about how things are in the actual world. The basic intuition is just that phenomenal properties are not identical to material properties. That is, the intuition is one of actual non-identity, not some fancy modal intuition. We simply refuse to identify phenomenal properties with material properties, even given all the arguments, in the way that we readily accept that Cicero=Tully.

Of course, given that we continue to remain convinced of non-identity in this way, even in the face of the arguments, we will also continue to think that zombies and ghosts are possible, that is, that the correlation between phenomenal and mental properties found in the actual world could be different in other possible worlds. But note that, on this line of thought, we continue to believe that this correlation is contingent as a result of continuing to believe that phenomenal properties are not identical to material ones, and not the other way round.

So my explanation of why we continue to believe in the possibility of zombies and ghosts, even after seeing the arguments, proceeds in a direction opposite to that suggested by the Kripkean analysis. It is not that we first feel some modal intuition, that zombies and ghosts are possible, and then infer that phenomenal and material properties cannot actually be identical. While this would be a valid enough inference, it is not the one I take to be responsible for our continuing resistance to materialism. Rather, we start at the other end of this inference, with the basic intuitive judgement that mind and brain are actually distinct, and then infer, again validly enough, that zombies and ghosts are possible.
The Phenomenon of Consciousness
David Papineau

Chapter 4

Phenomenal Concepts

4.1 Introduction

I have argued that materialists should be conceptual dualists. At the ontological level, of course, they must be monists, identifying phenomenal properties with material properties. But at the level of concepts they should distinguish two different modes of reference to these phenomenal/material properties. In addition to the possibility of referring to these properties as material, they should also recognize a distinct mode of referring to those properties, via phenomenal concepts which pick out those properties, so to speak, in terms of the way they feel.

The initial reason for recognising phenomenal concepts was Jackson's knowledge argument. While this argument failed to demonstrate that phenomenal properties are non-material, it did at least show that there are distinctive ways of referring to those properties, ways that are standardly only available to human beings after they have had some original instance of the experience in question. Kripke's modal argument then told us something further about phenomenal concepts. They must somehow refer directly, not via description. They don't identify their referents as the bearers of some further property they may contingently possess.

So far, then, we know two things about phenomenal concepts. Their possession is standardly consequent upon some earlier version of the type of experience they refer to. And they refer to that experience directly, and not via some description. In this chapter I want to build on this basis to develop a more detailed understanding of the structure and referential power of phenomenal concepts.

4.2 Psychological, Phenomenal, and Everyday Concepts

It will be helpful to start by clarifying the relationship between everyday thought and the conceptual dualism I have been urging. I want to distinguish between material and phenomenal concepts of experiences. However, this distinction plays no prominent role in everyday thought. Everyday discourse uses undifferentiated words for phenomenal states, like "pain", or "hearing middle C", and does not stop to specify whether these words should be understood as expressing phenomenal concepts or material ones.

I think we should view everyday terms like "pain" and "hearing middle C" as simultaneously expressing both sorts of concepts. Before considering exactly how this expressional duality might work, let me be a bit more specific about the kinds of material concepts that might plausibly be expressed, along with phenomenal concepts, by everyday discourse. The relevant concepts here will standardly be concepts associated with causal roles, concepts that pick out their referents in terms of a structure of macroscopic causes and effects (such as bodily damage and avoidance
behaviour in the case of pain, or ambient sounds and musical responses in the case of hearing middle C). While physiological experts like Mary may often know a lot more about the physical realization of such causal structures in specific kinds of beings, I take it that such detailed physical information is not normally part of pre-theoretical everyday thought. So, insofar as everyday thought does utilise non-phenomenal physical concepts of experiential properties, these will be concepts involving everyday causal roles, rather than concepts that pick out their referents in terms of specific physical constitutions. Following David Chalmers (1996), I shall call these "psychological concepts".

So the thought I wish to pursue is that an everyday term like "pain" expresses both a phenomenal concept of pain, the concept of a state that feels a certain way, so to speak, and a psychological concept of pain, the concept of a certain causal role. Does this mean that the everyday term "pain" is equivocal, expressing two quite different ideas, which careful users of the language need always to disambiguate? Not really. The situation here is different from the paradigm case of ambiguity, where a given syntactic form ("bank", or "bat")expresses two quite unrelated concepts which refer to two quite different entities. For the presumption of everyday thought is surely that the two concepts, the phenomenal and the psychological concept of pain, are connected by the fact that they actually refer to the same state.

To see how this might work, consider the case of multi-criterial concepts, of the kind often found in science, where two independent criteria (resistance to acceleration and gravitational charge, say) are both regarded as diagnostic of some kind (mass). Such multi-criterial concepts are likely to display a species of semantic indeterminacy, in that in many such cases it will be left indeterminate how exactly the various criteria fix the referent. For a body to have a given mass \( m \), must that number measure both the body's resistance to acceleration and its gravitational charge? Or is it enough that it measure resistance to acceleration, or, alternatively, that it measure gravitational charge?

This kind of semantic indeterminacy is significantly different from that displayed by "bank" and "bat". This is because in most such cases the different possible criteria will all in fact fix the same referent, and so scientists who take the criteria so to work in concert will see no need to do any semantic refining. Newtonian physicists never felt obliged to decide between resistance to acceleration and gravitational charge as criteria for mass, precisely because they believed that the same quantity would be picked out either way. (Cf. Papineau, 1996.)

Similarly, I say, with everyday discourse and "pain". The term "pain" does indeed express two conceptually independent notions, phenomenal and psychological. But since it is generally assumed that these two concepts refer to the same state, everyday thought does not exert itself to decide which concept the term "pain" really expresses.

Of course, semantic refinement of a multi-criterial concept can become mandatory, if new discoveries overturn the empirical assumption that the different possible uses of the criteria all pick out the same kind. When general relativity showed scientists that resistance to acceleration and gravitational charge can come apart, it was no longer possible to work with the old unrefined Newtonian concept of mass, and scientists were forced to distinguish inertial mass from rest mass.
Similarly, new theories of conscious states could force everyday thought to refine its bivalent usage of terms like "pain". Interestingly, neither materialism nor interactionist dualism puts any pressure on everyday usage, since both of them agree that the phenomenal and psychological concepts of mental states pick out the same entities. But epiphenomenalism (or psychophysical parallelism) would call for some refinement of everyday usage, since on these views phenomenal and psychological concepts will no longer co-refer. So then a decision would be needed. Thus, if we embraced epiphenomenalism, should we conclude that nothing satisfies the word "pain", on the grounds that nothing fits both the phenomenal concept and the causal role concept of pain? Or should we take "pain" to refer to the inefficacious epiphenomenal state, which presumably satisfies the phenomenal concept, but not the causal role one—or, alternatively, to the efficacious physical state, which satisfies the causal role but not the phenomenal concept?

Perhaps the middle of these options strikes you as most plausible—epiphenomenalists should aim to use "pain" to refer to their putative inefficacious epiphenomenal states. But, in any case, there is no substantial issue here. It is simply a matter of how the word "pain" should be used by people who don't think that phenomenal and causal roles concepts co-refer. The rest of us, who think these concepts do co-refer, can carry on using the word unreflectively, as picking out that common referent.

Of course, none of this denies that there are two concepts here, which both refer in their own right, and indeed each of which could be expressed by a different word ("psychological-pain" vs "phenomenal-pain", perhaps). My discussion of the everyday word "pain" has simply aimed to show that there is no reason why everyday usage should have decided between these as the conceptual content of the word "pain" we actually use.

My focus in this chapter will now be on the phenomenal concept of pain. The idea is thus that we should peel off a purely phenomenal element from the notions expressed by the everyday term "pain". This will be what we are left with, so to speak, when we have subtracted all psychological ideas of pain, all ideas of pains as things with certain characteristic causes and effects. Our task is to understand how this purely phenomenal concept is structured, and in virtue of what it has its referential power.

At some point some readers may be becoming uneasy. Could such a purely phenomenal concept really succeed in referring at all? Is not this idea dangerously close to a private language for mental "givens"? I shall not address such worries directly. But by the end of the chapter I hope that worried readers will feel easier on this issue. The account of phenomenal concepts I develop will have space for a certain kind of non-phenomenal element, and this will help reduce anxieties about privacy. But that is for later. Having flagged this issue, I shall now ignore it until the end of the chapter.

Let me conclude this section by drawing attention to a feature of phenomenal concepts which has so far been left implicit, but which will prove of some interest in this chapter. This is the fact that phenomenal concepts can refer both to particular experiences and to types of experience. This ability, to refer to both particulars and to types, is displayed by other types of general concepts. (Thus we can say both that
"The electron is attached to the oil drop" and that "The electron has negative charge"; or, again, "The whale has escaped" or "The whale is a mammal".) Phenomenal concepts are similar in this respect. Thus Mary might use imaginative recreation to think about a type, as in, "That experience was very exciting—I hope I have it again"—or, alternatively, to think about a particular experience, as in, "That experience must have been caused by what I ate last night". And the same contrast will be present in thoughts grounded in introspective classification. Thus "I wouldn't mind having this experience more often" versus "This experience can't last much longer".

4.3 Indexicality and Phenomenal Concepts

Some of the phrases I have been using to express phenomenal concepts may have made it seem as if phenomenal concepts are simply a species of indexical concept. I have typically alluded to imaginative uses of such concepts with the construction "that experience", and to introspective uses with the phrase "this experience". This might suggest that there is nothing more to phenomenal concepts that the possibility of referring to experiences using standard indexical constructions.

Moreover, this idea might at first sight seem to fit with what we already know about phenomenal concepts. Maybe the reason why Jackson's Mary does not have a phenomenal concept of seeing red prior to having the experience itself is simply that previously she didn't have any sample of red experience to refer to demonstratively, while she gains such a sample with her new experience. And similarly, in connection with Kripke, the indexical idea promises to explain how phenomenal concepts can refer distinctively without invoking any distinctive property of their referent: they refer indexically, by simply pointing, as it were, to instances of the experience that stand in some specifiable relation to the speaker.

However, I don't think this indexical idea works. At the end of this chapter, I shall return to this topic, and qualify this denial in certain ways. But at this stage we can show that indexicality on its own is certainly not sufficient to account for the workings of phenomenal concepts. We cannot explain Mary's new conceptual abilities simply in terms of her now being able to apply old indexical constructions, of a kind that were always available to her, in her new post-experiential context.

As a preliminary to showing why, let me first sketch a crude model of indexical constructions in general. Let us take it that indexical terms always involve a demonstrative element ("this", "that", or perhaps simply pointing) plus a descriptive element ("animal", "shape", "car"). The compound indexical term ("that animal") then refers to the unique entity, if there is one such, that both lies in the "direction" indicated by the demonstrative element and satisfies the descriptive term. Some indexical phrases run together both demonstrative and descriptive components ("now" = "this time", "there" = "that place"), but this terminological fact does not affect the underlying model.

We can think of the descriptive element in an indexical construction as fixing some range of possible referents, and the demonstrative element plus the contents of the indicated "direction" as then narrowing down this range to some specific referent.
Note that it is consistent with this model that indexical constructions—just like phenomenal concepts—can be used to refer both to types and to particulars. The phrase "that car" can be used both to pick out a model (the Rolls-Royce Corniche, say) or some specific car (Tom Jones's Roller). The disambiguation here can be done explicitly ("that make of car") or left to conversational context.

Let me now return to phenomenal concepts themselves. The suggestion I want to examine is that the distinctive nature of phenomenal concepts can be fully explained by viewing them as normal indexical constructions. Given the model of indexical constructions just outlined, we can take this to mean that there are no distinctive demonstrative or descriptive constructions possessed by users of phenomenal concepts. Mary has just the same demonstrative and descriptive resources as she had before she left her house. Rather, her distinctive new powers of reference lie solely in the fact that she in now in a position to demonstrate experiential items she previously could not.

The sharpest way of showing that this suggestion does not work is to consider Mary's ability to think imaginatively about red experiences after her original red experience is over. She is able to re-enact her original experience in imagination, and thereby to to think about that experience (that type, let us suppose, for the sake of specificity). Now, as we have seen, it is natural to represent her referring term as "that experience". But it easy to see that there must be more involved here than Mary's now being able to demonstrate the type in question by pointing back in time to one of her own experiences.

For note that Mary may have lost track of when and where she had this earlier experience, and so may not be able to identify its spatio-temporal location using standard demonstrative constructions. This won't stop her being able to refer to that type of experience as "that experience", accompanied by some act of imagination. She will still be able to think thoughts like "I'd love to have that experience again, even though I can't for the life of me remember when or where I previously had it."

This example shows that imaginative references to types of experience must depend on something more than ordinary indexical constructions. Mary can't simply be pointing back, using standard indexical devices, to the original occasion where she first had that kind of experience. If there is anything indexical here, it must involve some special referential power associated with the imaginative act itself. In section 4.10 below I shall explore this possibility a bit further. But for now let us simply note that any such special imaginative indexicality will depend on something more that indexical devices which were always available to Mary. For it will make essential use of Mary's ability to imagine seeing red, and this is something she acquires only after her original red experience.

It may seem as if the indexical model will work better when we turn to introspective uses of phenomenal concepts, rather than imaginative ones. When I look into myself, and refer to some aspect of my experience as "this feeling", what am I doing, except pointing internally to something occurring inside me?

But even here there are difficulties. For a start, we might query the equation of "internal pointing" with some standard spatio-temporal demonstrative construction like
"inside my body now". Moreover, even if we leave this worry to one side, there is a question about the descriptive concepts which are being used to facilitate the introspective reference to experience. At any time an individual's conscious experience will be multi-faceted and multi-modal. You can see many different features and objects at a given time, not to mention further awareness involving hearing, smelling, itching, and so on. Given this, you will need something akin to a descriptive concept to determine which aspect of your current experience is being referred to. However, if the items which play this descriptive role are only available to Mary after her original experience, then this will again discredit the indexical hypothesis we are examining, since that hypothesis assumes only referential constructions that were always available to Mary.

It may seem as if material concepts of experience could fill the bill of descriptive concepts that were always available to Mary. After all, even before she left her house, Mary had a full range of material concepts of experience. She was thus able conceptually to distinguish, in material terms, between colour experiences, brightness experiences, and so on; and also between vision as such, as opposed to hearing, smelling, and so on. So why can't she use these material concepts to identify whichever aspect of her current experience she wishes to refer to, as in "this colour experience" or even "this smell"? On this model, the actual contents of her current consciousness will still play a role in determining what is being referred to. Thus, the internal pointing will direct us inwards towards consciousness, the material description ("colour experience") will specify what kind of experience is at issue—and current consciousness itself will then determine that it is an experience of green, say, rather than red or yellow, that is actually designated.

However, this kind of use of material concepts cannot account for introspective reference. To see why, we can use a variant of the Mary thought experiment. Suppose Mary's sensory deprivation has been rather different. Mary has never heard anything or smelled anything. Then she has a new experience. She hears middle C, say. Suppose we set things up so that she doesn't know, in her old material terms, whether this is a sound or a smell. Now, I take it that this won't stop her being able to refer introspectively to this or any further such experiences as "this experience", picking them out from the other aspects of her overall experiential state with the help of a phenomenal concept she has now acquired.

But, if this is right, this use of her phenomenal concept cannot be equated with something along the lines of "this" (pointing internally) plus a descriptive concept drawn from her old repertoire of material concepts. She will have material concepts like "hearing a noise", "smelling", and so on. But she won't know which of these to use to form the relevant indexical term. Since she won't even know, in her old material terms, whether the experience she is introspectively classifying is a smell or a sound, she won't know which of those old material concepts would serve her indexical purpose.

So, as with imaginative uses of phenomenal concepts, it seems as if something derived from prior first-hand experiences plays a crucial role in determining what introspective uses of phenomenal concepts refer to. We have yet to understand exactly how this
works. But we can be sure that it involves something more than indexical complexes built up from elements that are available prior to first-hand experience.\footnote{For some rather different criticisms of an indexical account of phenomenal concepts, see Tye, 1999.}

4.4 Phenomenal Properties Provide Their Own "Modes of Presentation"

In this section I want to cast doubt on a different suggestion about phenomenal concepts. One sometimes sees it said that phenomenal properties can provide their own "modes of presentation" (cf. Loar, 1999). This thought is often associated with the claim, defended in the last chapter, that phenomenal concepts refer to phenomenal properties directly, and not by invoking any further contingent properties of those referents. While I of course agree with this latter claim, I think that the frequently accompanying talk of "modes of presentation" needs to be treated with extreme care.

At one level, the idea that phenomenal properties can provide their "own modes of presentation" may simply mean that they do not have to be picked out via some other contingently connected property they possess. Here there is no dispute. There is only one property in play when a phenomenal concept refers, namely, the phenomenal/material referent itself. No further property mediates between referring concept and referent.

So far so good. But sometimes something more seems to be meant, and here I think we need to be careful. I take it that a "mode of presentation" is something like a Fregean sense, something grasped by the mind and with some kind of semantic power to latch onto a referent. The paradigm, perhaps, is where the mind is already able to think of some property, or combination of properties, Ø, and then uses this ability to construct a term to refer to the entity which possesses those properties ("the thing which has property Ø").

Now this Fregean picture of "modes of presentation", I take it, indicates that we ought not to talk about phenomenal properties providing their "own modes of presentation". The idea we are working with is that phenomenal concepts refer to phenomenal properties directly, without mediation of any further properties. It would seem badly to misrepresent this idea to say that phenomenal properties provide their own Fregean "modes of presentation". This suggests a picture whereby the mind somehow already has the power to think about some phenomenal property, Ø, and then uses this ability to form a mode of presenting that property ("the property which is property Ø", perhaps). But this makes little sense. If we already have the ability to think about the phenomenal property Ø, we don't need to construct some further mode of presentation to enable us to think about it.

But there is a further circumstance in the offing which is capable of obscuring this point, and indeed of sowing great confusion about consciousness generally. Note that when we deploy phenomenal concepts, we will also characteristically instantiate some verison of the conscious property we are referring to.

This is most obvious with introspective uses of phenomenal concepts. When I pick out some aspect of my current experience introspectively ("this feeling . . ."), I will be
having that feeling at the same time as referring to it. And a similar point applies to imaginative uses of phenomenal concepts. When I later think imaginatively about some earlier experience, like seeing red ("that experience . . ."), I won't actually have an experience of seeing red itself, but this experience is likely to bear some phenomenal similarity to the experience of seeing red—to be "a faint copy", as Hume put it.

So in both cases the use of phenomenal concepts to refer to some experience will standardly involve the thinker actually having the experience itself, or a faint copy of it. Perhaps, though this is yet to be determined, we should think of this instantiation of the experience as literally part of the term the thinker uses to refer to that experience. And, even if we don't go that far, we should certainly recognize that uses of phenomenal concepts will standardly be accompanied by versions of the experiences referred to.

Now, I take this feature of phenomenal concepts to be hugely important. Indeed in two chapters time it will provide the crucial ingredient for my explanation of "the intuition of distinctness"—that is, the widespread and well-nigh inescapable feeling that conscious and material properties must really be distinct. To give a quick preview, in chapter 6 I shall argue that this intuition cannot be adequately explained just by the fact that we have two modes of referring to conscious/material properties, with the phenomenal mode referring directly. Nor is it adequately explained by the further fact that certain kinds of explanation of the relation between physical and conscious properties are thereby ruled out (this "explanatory gap" will be analysed next, in chapter 5.) So in chapter 6 I shall give a different explanation of the insistent anti-materialist intuition, which rests specifically on the fact that uses of phenomenal concepts are necessarily accompanied by versions of the conscious experiences being referred to. More specifically, I shall argue that this feature of phenomenal concepts leads us, first, to think that material concepts of conscious experience will inevitably "leave out" the experiences themselves, and that this then seduces us into the fallacious conclusion that material concepts cannot refer to conscious experiences.

But all that is for later. Our current concern is not with confusions that might be generated by phenomenal concepts, but with the analysis of how these concepts work in the first place. And here I think that the fact that uses of phenomenal concepts involve versions of their conscious referents is of no immediate importance. In particular, I do not think that this fact generates any immediate explanation of how those concepts refer to those experiences.

It is possible, however, to construe the idea that "phenomenal properties are their own modes of presentation" as offering just such an explanation. The thought here would be that, in deploying phenomenal concepts, the mind is somehow in possession of an instance, or version, of the property being referred to, and that this in itself immediately accounts for the fact those concepts refer to those properties.

I think this thought must be resisted. It betrays loose thinking about reference to suppose that concepts will automatically refer to any properties that are involved in their deployment. Maybe the involvement of conscious properties in phenomenal concepts will turn out to be of some significance to those concepts' referential powers. But this can't be the whole story. After all, entities don't normally refer to themselves. So why should the presence of a conscious property in the mind automatically
constitute a term which refers to that property? Still less do entities normally refer to whatever they might be "faint copies" of. So, again, merely instantiating a faint copy in imagination will not automatically constitute a term which refers to the original of that faint copy.

4.5 World-Directed Perceptual Recollection and Classification

At this stage it will be helpful to turn away from phenomenal concepts, and consider some closely related mental powers.

Let me start by expanding my treatment of sensory imagination. In my discussion so far, this has figured in what I have been calling "imaginative uses of phenomenal concepts". But a moment's reflection will show that this kind of contribution to thoughts about experiences is not the only way—nor indeed the most basic way—in which sensory imagination can contribute to thought. For we can also use sensory imagination to think about non-mental things, like trees, or houses, or other perceivable objects.

For example, when I visually imagine the beach next to the house where I grew up, in Isipingo, South Africa, I do not normally do so in order to think about the visual experiences I used to have, but in order to think about the reefs, sandbanks, waves, and rockpools I so enjoyed. Again, when I visually recall entering the Departmental Office in King's College London this morning, the normal upshot is that I think about the contents of the room, like desks, computers, and the departmental administrator, not about my matitudinal visual experiences.

My point here is that sensory imagination is in the first instance a medium for thinking about the external world of macroscopic physical objects, and only secondarily a means of thinking about experiences themselves. To drive the point home, we need only consider the possibility of thinkers who are incapable of thinking about experiences as such, who have no notion of mental states. It seems clear that such thinkers could still use sensory imagination to think about the world, to think about rockpools or desks, say, even though they can't use imagination to think about experiences. So parallel to—indeed prior to—the use of sensory imagination to think about experiences, there is a more basic use of sensory imagination, to think about ordinary non-mental things.

Now, a quite analogous point can also be made about the other use of phenomenal concepts, in introspective classification, as when we focus on some aspect of our current experience, and think "this feeling . . .", "this colour experience . . .", and so on. Parallel to this kind of introspective classification of experience stands ordinary perceptual classification. When I am looking at a visual scene, I will visually classify certain aspects of that scene. For example, when looking out to sea, I will see the waves as waves, say, and the seagulls as seagulls. Or in the office I might identify the new i-Mac as such, perhaps as a result of noting that it has one of those curious colours.

Again, perceptual classification seems prior to phenomenal classification. While some of the same sensory powers may be involved, their basic use is surely to think about
the external world, rather than about experiences themselves. Consider again thinkers who are incapable of thinking about mental states as such. They can still use their powers of sensory classification to think about the world, to think about waves and computers, even though they can't use them to think about experiences.

4.6 Perceptual Concepts

In two sections I shall consider the relationship between perceptual thinking about the non-mental world, on the one hand, and phenomenal thinking about experiences, on the other. But first let me say some more about the former world-directed powers.

To help keep things clear, I shall henceforth use the terms "perceptual recall" and "perceptual classification" specifically to refer to world-directed acts of imagination and classification respectively, and I shall also talk about these as two uses of "perceptual concepts". When I want to talk about the corresponding uses of phenomenal concepts to refer to experiences, I shall continue to speak of "imaginative recreation" and "introspective classification".

Note, to start with, that there is a question about whether we should talk about separate recollective and classificatory perceptual concepts, as opposed to counting these as two uses of single perceptual concepts. This parallels the corresponding question which came up in connection with phenomenal concepts at the end of chapter 2.

Thus, consider my perceptual concept of "that kind of bird", where I don't know anything else about the kind of bird in question, but can classify it visually, and can recall it in visual imagination. Now, the classificatory power involved here seems dissociable from the recollective power, and vice versa. It is easy enough to think of cases where one can classify something perceptually when it is present, but cannot recall it in perceptual imagination. And nothing a priori seems to rule out the possibility of someone who can recall something perceptually, even though they are no good at picking it out when it it present (though this would admittedly be somewhat stranger). Given this possibility of dissociation, should we not rather recognize two different kinds of perceptual concept, recollective versus classificatory, rather than one kind of concept variously deployed?

However, as with the corresponding question about perceptual concepts, I shall not fuss about this issue. Once more, it is a matter of description rather than substance. So sometimes I shall talk about perceptual concepts simpliciter, and at other times I shall distinguish between recollective and classificatory uses of these concepts.

Another way in which perceptual concepts are like phenomenal concepts is in their dependence on prior experience. Possessing a perceptual concept of some entity normally requires that you have previously perceived that entity. You will not be able to classify something visually as a certain kind of bird, say, or as a certain colour, unless you have seen it before, nor will you be able to think about it using perceptual recall. This mirrors the point that the possession of a phenomenal concept requires that you have previously undergone the experience that concept refers to.
Of course, the normal qualification is needed here, to allow that we can think perceptually about complex objects—turquoise squares, say—that we have never seen before (provided, that is, that the requisite simple concepts have been derived from previous perceptions of turquoise things, and square things). And a further qualification is needed, in the case of perceptual concepts, which doesn't apply phenomenal concepts. For you can acquire a perceptual concept of a kestrel, say, even though you haven't perceived any instances, provided you have seen something which produces the same perceptual reactions—such as a picture of a kestrel, say, or a video of a kestrel in flight.

No doubt the explanation here is the same as for the corresponding point about phenomenal concepts. When we deploy a perceptual concept to think about some non-mental entity, we will be activating some neural pattern. However, an original perception will have been needed to fix that pattern as something that can be so activated. As before, the brain needs an original from which to form the mould for further activations.

4.7 How do Perceptual Concepts Refer?

Let me now focus on the referential powers of perceptual concepts. What makes it the case that my perceptual concept of "that kind of bird" indeed refers to the kind of bird in question? Clarity on this issue will bear fruit when we turn, in the next section, to the corresponding question about the referential powers of phenomenal concepts.

A first thought might be that perceptual concepts refer in virtue of the fact that they resemble their referents. I assume that this suggestion does not need to be taken seriously. It is true, to stick to the same example, that the bird in question will "look" as things appear when we exercise visual concepts of "that kind of bird". (And, similarly, it will "sound" as things seem when we exercise aural concepts of it, and "smell" as things seem when we exercise olfactory concepts of it, ...). But this is just the definitional truism that how the bird "looks" to us is a matter of how we normally respond to it visually. To explain why those responses are about the bird in the first place would seem to require some more basic resemblance, between the bird itself, in abstraction from how it appears visually, and exercises of our visual concept of it. I know of no good way to make sense of this idea.

A second thought might be that perceptual concepts refer via descriptions which invoke phenomenal properties. Thus thoughts involving the perceptual concept "that kind of bird" might be contrued as equivalent to "the kind of bird which produces these visual experiences". And in general perceptual concepts could be analysed as equivalent to "the which produces sensory experiences Ø". This would then make the referential powers of perceptual concepts derivative from those of phenomenal concepts. Of course, these latter referential powers, of phenomenal concepts themselves, have yet to be explained. But the idea would be that, however they are explained, the referential powers of perceptual concepts would piggy-pack on them, in virtue of descriptive definitions along the above lines.

I have a simple objection to this second thought. The referential powers of perceptual concepts must be independent of those of phenomenal concepts. For surely it is
possible for there to be beings who possess perceptual concepts, but no corresponding phenomenal concepts. These would be beings who are capable of thinking in perceptual terms about birds, trees, shapes, and colours—indeed about anything perceptible—but who have no concepts of sensory experiences, or of perceptions, or of minds generally. I shall say rather more about this possibility in the next section. But I take it to be relatively uncontentious that some of our evolutionary ancestors must have been like this, if not some existing higher animals; moreover, it seems likely that many autistic people are also like this.

If this is right, then perceptual concepts cannot get their semantic power from descriptions framed using phenomenal concepts. For there will be beings who can deploy the perceptual concept "that kind of bird", say, yet are incapable of thinking about the visual experiences characteristically produced by that bird. So their perceptual concept must derive its referential powers from something other than an association with a description which uses phenomenal concepts.

Perhaps this is a bit quick. Once modern humans are sophisticated enough to possess phenomenal concepts, then they will certainly be capable of forming descriptions of the form "the which produces sensory experiences Ø". And so perhaps they will use such descriptions to replace, or transform, some of the simpler perceptual concepts deployed by their less sophisticated ancestors and other beings. Maybe this is what happens with concepts of "secondary qualities", among people who are reflective enough to find reason to distinguish such qualities from other features of the external world.

Still, I find it quite implausible to suppose that all the perceptual concepts of normal modern humans have been so transformed. Apart from anything else, the possibility of beings who possess perceptual concepts, but no phenomenal concepts, shows that perceptual concepts are capable of referential powers in their own right, independently of any association with descriptions involving phenomenal concepts. Given this, it is hard to see what motivation there could be for transforming all perceptual concepts into phenomenal-concept-involving descriptions. Maybe such transformations are justified in special cases, such as by the kind of considerations that might motivate demarcations of "secondary qualities". But this is no reason not to use untransformed perceptual concepts in their own right when such considerations do not apply.

This now returns us to the problem of explaining the referential powers of untransformed perceptual concepts. If they do not refer via descriptions relating them to phenomenal concepts, how exactly do they refer? I think the way forward here is to appeal to naturalistic theories of representation, in the style of causal or teleosemantic theories.

The simplest version of such a theory would be a straightforward causal account, which held that a perceptual concept refers to that entity which normally causes classificatory uses of that concept. For example, a perceptual concept might refer to some kind of bird because it is specifically birds of that kind which cause classificatory deployments of that concept.

The difficulties facing this simple causal story are well-known. Most centrally, classificatory deployments of perceptual concept are often caused by things which the
concept doesn’t refer to. You can be fooled into visually judging that some kind of
bird is present by mechanical birds, pictures, or tricks of the light. Yet your concept
doesn’t refer to a kind which includes these deceptive stimuli. This refutes the simple
causal story. The trouble, in effect, is that the simple causal account of representation
leaves no room for misrepresentation.

Teleosemantic theories deal better with misrepresentation. Theories of this kind ask
about the purpose of the perceptual concept, in a biological sense, rather than it
causes. The referential value of the concept can then be equated with those items
which it is the biological function of the concept to track. Since concepts can
malfunction, like other biological traits, it no longer follows that misrepresentation is
impossible. Sometimes a concept will be activated when it is not supposed to be. (Cf.
Millikan, 1984, 1989; Papineau, 1984, 1993a.)

This is not the place to pursue details. In what follows, I shall simply assume that
referential powers of perceptual concepts can be explained by some version of
teleosemantics, or, perhaps, by some revision of the causal theory sophisticated
enough to deal with misrepresentation (cf. Fodor, 1990). Fortunately, none of the
arguments which follow will depend on the exact form of such a naturalistic version of
representation. (It will be of some significance, however, that such naturalistic
theories will portray perceptual concepts as referring directly, in the sense that they
have referential powers which do not derive from their association with any further
descriptions.)

4.8 The Phenomenal Co-Option of Perceptual Concepts

I turn now to the referential powers of phenomenal concepts. Even if we assume that
perceptual reference can be explained naturalistically, this does not yet tell us about
phenomenal reference. We still need to explain the ability of phenomenal concepts to
refer to phenomenal properties.

I take the following to be the obvious way of understanding phenomenal concepts.
Originally there were just perceptual concepts. Our distant intellectual ancestors could
perceptually classify things as birds, faces, colours, and so on. Moreover, they could
use their powers of perceptual recall to think about such things even when they were
absent. But they couldn't think about experiences.

Then they built on this basis to construct a practice for thinking about experiences
themselves. An natural hypothesis is that they started to deploy concepts of the form
"the experience: ———", where the gap was filled by some actual act of perceptual
classification or perceptual recall. By prefixing this act with the operator "the
experience: ———", they were able to generate terms which referred to the
experiences themselves. Thus, for example, you might visually imagine something red,
and by prefixing this act with the experience operator, form a term apt to refer to red
visual experience, as opposed to referring to a red surface. Or you might do the same
while perceptually classifying some object as red, and again form a term with
equivalent referential content, a term which refers to your experience of seeing red,
rather than to a red surface.
It is plausible to regard the availability of these terms as part of the emergence of "understanding of mind". Human beings have a highly developed facility for thinking about their own and other individuals' mental states. The classic manifestation of this is their success at the "false belief test", which requires the attribution of mistaken representations to other agents. Human children are able to do this from the age of three or four onwards, though not before. It is unclear whether other animals can reason about minds to a similar extent.

Most discussions of "understanding of mind", in this sense, have focused on the ability of humans to attribute beliefs and desires to each other, and to use these attributions to predict behaviour. In particular, there has been a detailed debate about whether we generate these predictions by simulating the decisions we would make if we ourselves had those beliefs and desires (the "simulation-theory"), or whether we deduce the predictions from some general theory of the way beliefs and desires cause actions (the "theory-theory"). (Cf. Davies and Stone, 1995a, 1995b; Carruthers and Smith, 1996.)

We need not enter into these issues here. For a start, my current interest in "understanding of mind" is somewhat different from the standard one, in that I am concerned with the ability to think about conscious experiences, rather than about beliefs and desires. No doubt there will be some connections between these kinds of conceptual powers. In particular, it would be surprising to find thoughts about beliefs and desires in the absence of thoughts about experiences. Still, I have no special views about the way we refer to beliefs and desires, and so can leave controversies on this matter to one side. My concern is solely to understand terms for phenomenal experiences.

Moreover, I have no need to take sides in the dispute between "simulation-theory" and "theory-theory". In one sense, it is true, I have suggested that phenomenal thinking about experience involves a kind of simulation: I have hypothesised that mental terms for experiences are formed by adjoining an "experience operator" ("the experience: — ——") to an actual act of perceptual imagination or perceptual classification. In this respect, I do think that phenomenal thoughts about experience involve a kind of simulation or instantiation of the experience being thought about. But this involves no commitment to the distinctive claims of "simulation-theory". It does not follow that any behavioural predictions drawn from such phenomenal thoughts must be generated by "off-line" simulations of the way those experiences might lead to decisions. Even if you form terms for conscious experiences by activating some version of the experience itself, you may still reason with the terms so formed in a quite theory-driven manner.

4.9 The Experience Operator

Let me now make some comments about the "experience operator" used to transform exercises of perceptual concepts into exercises of phenomenal concepts. I shall discuss the contribution of the perceptual concepts themselves to the compound phenomenal terms in the next section.

An initial question is whether we should think of the experience operator as incorporating some ordinary non-phenomenal notion of experience. This would be a
notion of experience as a psychological type, as a state with certain causes and effects, rather than of experience as the type of state that it is "like something" to have.

Of course, such a generalized psychological notion of an experience-as-such, of the determinable experience, rather than of any specific determinate like seeing a kestrel, would not allude to any specific causes and effects. My thought here is rather of a general psychological notion of a state with some—as yet unspecified—non-mental (and perhaps mental) causes, and some—as yet unspecified—non-mental (and perhaps mental) effects.

The alternative would be to take the notion of experience involved in the experience operator as some sui generis phenomenal concept, perhaps some primitive notion of a state that is "like something". Such a concept would have no a priori links to any non-phenomenal psychological concepts alluding to causal roles. If we take this option, then we will need to explain the semantic powers of the experience operator directly, rather than as deriving from the meaning of some notion of experience which also plays a role in non-phenomenal thought.

This seems to me to be a possible line. There is no obvious reason why we should not explain the semantic significance of the experience operator directly. Such an explanation would presumably appeal to the contribution of the experience operator to the causal or teleosemantic significance of compound concepts formulated using it.

Nevertheless, it seems to me more attractive to view the experience operator as making use of some generalized psychological concept of experience, some concept of a range of psychological state-types which can be identified by their distinctive—though as yet unspecified—causal roles. The advantage of this strategy is that it builds it into phenomenal thinking about experiences that they do have characteristic—even if initially unknown—causal roles. I take this to be an advantage because there would be little point to phenomenal thinking which was not able to identify the phenomenal states it referred to as bearers of causal roles.

There is room for debate about the primary function of phenomenal thinking. Did it first emerge in order to enable our ancestors to predict the behaviour of others, or to anticipate their own future experiences, or even to reflect on the epistemological credentials of their own and others' beliefs? Still, it is clear that on any one of these stories, or indeed on any similar story, our ancestors would have needed to know something about the characteristic causes and effects of phenomenal states (in order to be able to infer what they will make people do, or when they will arise, or how they can mislead people as to the facts).

Of course, it is definitive of determinate phenomenal concepts, as I have introduced them, that they do not identify their referents a priori as bearers of certain causal roles. So I am not now suggesting that phenomenal concepts formed via the construction "the experience: ———" will be associated a priori with some causal role distinctive of the state being referred to. The whole point of phenomenal concepts is that you can have them and yet be quite ignorant of the distinctive causes and effects of the state referred to. (Mary can think "the experience: ———", recalling what first happened to her when she left her house, and yet have no idea of the characteristic causal role of that state.)
My present point is only that it may still be determined a priori that the referents of phenomenal concepts have some characteristic causes and effects, even if those phenomenal concepts per se do not determine a priori what those causes and effects are. And this much I take to follow from the suggestion that the "experience operator" utilises some generalized psychological concept of experience. When Mary thinks "the experience: ——— . . .", she knows a priori that, if she is thinking of anything, she is thinking of some psychological state with distinctive causes and effects. It will still be a matter of empirical research to figure out what those causes and effects are. But at least she will be given an initial push in the direction of such research by her a priori grasp of the experience operator.

This is scarcely a knockdown argument against the alternative idea that the experience operator involves a sui generis phenomenal concept, with no a priori links to any ideas about causal roles. For, even on this alternative, there is no absolute reason why thinkers should not be able to figure out that their resulting terms do refer to bearers of causal roles, perhaps using observations of brute empirical correlations as evidence for such identifications. However, there seems no good reason to suppose that our ancestors were faced with this testing intellectual task, especially given the point, made above, that any plausible function for phenomenal thinking will wait on some definite knowledge of the different cause and effects of different phenomenal states. It is far easier to understand the initial emergence of phenomenal concepts if we suppose that they incorporated an a priori pointer in the direction of such causal knowledge, as opposed to giving our ancestors no initial clue about this.

There is a further reason for thinking of the "experience operator" as involving psychological, causal-role ideas from outside the realm of phenomenal thought. Recall a point made in section 4.2, that we can use phenomenal concepts to refer both to particular experiences and to types of experience. A natural explanation is that these different uses derive from alternative versions of the experience operator. That is, we might form a term with the structure "the particular dated experience: ———"; alternatively, we might form a term with the structure "the type of experience: ——— ". In both cases the gap will be filled in with an actual exercise of perceptual attention or imagination. But the referential value of the two phrases will be quite different, as determined by the kind of operator involved.

Again, it may in principle be possible to explain such variant constructions directly, using causal or teleosemantic considerations. But it is surely far simpler to suppose that the variant constructions simply invoke a conceptual contrast available from other areas of thought, the contrast between the notion of a particular entity and the notion of a type.

At this stage, let me briefly return to a worry raised at the beginning of this chapter, about the apparent privacy of phenomenal concepts. The discussion in this section should have helped to allay such worries. For it shows that phenomenal concepts are not private in any worrying sense. While their deployment may involve distinctive experiential elements, it is no part of their semantics that they can only refer to the experiential states of the thinker who is deploying them. On the contrary, phenomenal concepts can be used refer to particular experiences in other people, and to types of experiences which are instantiated in many different people. Indeed I have suggested
that it may be built in to phenomenal concepts, via the "experience operator" involved, that their referents have characteristic causes and effects, of just the kind which might allow their attribution to other people.\(^\text{16}\)

### 4.10 The Perceptual Filling

On the picture sketched above, perceptual concepts are formed when we conjoin the "experience operator" with an actual act of perceptual classification or perceptual imagination. The idea, roughly, is that the concept so formed will refer to some experience that is associated with the classificatory or imaginative act involved. But I have as yet given no explicit account of the source of this referential power. Nor have I specified the precise relation between the experience referred to and the act of perceptual imagination or classification involved in the phenomenal concept.

One possibility at this point would be to regard this referential power as a relatively sui generis matter once more. On this conception, a given act of perceptual classification or imagination would make a systematic contribution to the referential value of compound concepts, but the nature of this contribution would be left to some naturalistic theory of reference to explain. Thus, the resulting compound concepts might be held to refer to those experiences which characteristically cause the concepts to be applied, as on a causal theory of reference, or to those experiences which the concepts are biologically supposed to track, as on a teleosemantic theory.

But perhaps a more informative story is possible. a story which does not leave the details of this referential relation to be determined entirely by empirical facts about the characteristic causes or functional effects of phenomenal thinking. I have in mind here a kind of quotational-indexical account of the workings of phenomenal concepts.

In section 4.3 above I argued that phenomenal concepts cannot be reduced to ordinary, non-phenomenal indexical constructions. Acts of imagination and classification of a kind that demand previous experience, play an essential role in securing phenomenal reference.

Even so, the model of phenomenal reference currently on the table does seem to admit of a different, yet still relatively straightforward, indexical reading. Why not simply read "the experience: ———", where the gap is filled by an act of perceptual classification or imagination, as a term which refers to the experience which fills the gap?\(^\text{17}\) This reading would automatically accommodate the earlier objections to an indexical story, since we are now explicitly viewing phenomenal concepts as

---

\(^{16}\) This is not to deny that there will be an asymmetry between the way phenomenal concepts are applied to other people, on the basis of behavioural and other evidence, and the way they are applied directly in introspection. Still, it can only be a virtue in a theory of phenomenal concepts that it implies such an asymmetry, together, ideally, with an account of the special authority of introspective uses. This is not the place to pursue this topic. But let me at least observe that it falls out of my overall story that introspective uses of phenomenal concepts require the presence of the experience referred to.

\(^{17}\) This model might seem to support a representational theory of consciousness. For it implies that the referents of phenomenal concepts are always acts of perceptual classification or imagination. But there are complications here, some relating to points I am about to make in this section, and some which will come up in chapters 7 and 8. Let me shelve this issue for the time being.
incorporating classificatory or imaginative acts, of a kind that require relevant previous experience. Given this, my new indexical suggestion might seem the obvious way to understand phenomenal reference.

On this picture, phenomenal concepts would work rather like quotation marks. The referring term incorporates the things referred to, and thereby forms a compound which refers to that thing. Thus, ordinary quotation marks can be viewed as forming a frame, which, when filled by a word, yields a term for that word. Similarly, my phenomenal concepts involve a frame, which I have represented as "the experience: ———"; and, when this frame is filled by an experience, the whole then refers to that experience.  

I like this quotational-indexical story. Apart from anything else, it has the virtue of explaining why it is so natural to represent phenomenal concepts with "this"s and "that"s. However, some significant qualifications are needed.

The most obvious is in connection with imaginative uses of phenomenal concepts. We are now assuming these have the form "the experience: ———", with the gap filled by an act of perceptual recollection—imaginatively recalling something red, say. And the suggestion just made is that the resulting term will then refer to the experience of imaginatively recalling something red. The trouble, however, is that the term in question does not standardly refer to this imaginative experience, but to the full-fledged experience it is a "faint copy" of. If I imagine a red square, say, and then think "that experience", I will not normally be thinking of the faint experience of imagining something red, but of the actual experience of seeing something red.

To cope with this difficulty, the obvious solution is to have imaginative uses of phenomenal concepts referring, not to the imaginative experience that is "quoted" itself, but to any experience that appropriately resembles it. To the extent that full-fledged experiences resemble the imaginative experiences that faintly copy them, this would then secure the desired reference to the full-fledged experiences.

This suggestion gains support from the model of linguistic quotation. When I use quotes to form a referring term, such as "antidisestablishmentarianism", I will not normally be using this term to refer only to the word as written in lower case and in this particular typeface. Instead I will be referring to a type which includes a wide range of possible inscriptions and sounds, with suitable linguistic or phonetic similarities to the exemplar within my quotation marks. Similarly, so the suggestion would go, with imaginative uses of phenomenal concepts. The phenomenal concept will refer to a type of experience which bears a certain resemblance to the "quoted" exemplar.

---

18 I am slightly hesitant about highlighting this analogy, given that quotation raises its own puzzles (cf. Cappelen and Lepore, 1997; Saka, 1998). These are therefore likely to arise for phenomenal concepts too. Still, I don't think these puzzles invalidate any of the further claims I make about phenomenal concepts.

19 A complication. Even if this is the normal case, note that we can also use an imaginative act to refer phenomenally to the faint imaginative experience itself, rather than to the full-fledged experience it faintly copies. I can perceptually recall seeing something red, think "the experience: ———", and thereby intend to refer to the faint recollective experience itself, rather than the actual experience of seeing something red. My hypothesis is that this involves some implicit modification of the "experience operator". My concept here is in effect "the imaginative experience: ———".
It is true that I am here helping myself to a notion of resemblance among experiences, a notion which has yet to be explained. I do not propose to pursue this topic, however. Perhaps the requisite notion can be viewed as piggy-backing on pre-existing judgements of perceptible resemblances among things perceived: experiences would then be counted as similar to the extent that the things they represented are perceptually similar. Or perhaps some other explanations of resemblances among experiences will work better. Rather than prolonging this already extended discussion, however, let me tie things up by briefly considering whether a quotational-indexical account of introspective uses of phenomenal concepts will also need to appeal to resemblances among experiences, in the way we have just seen is necessary for imaginative uses.

Here things are not so clear-cut. On the quotational-indexical account, introspective uses of phenomenal concepts fill the frame "the experience: ———" with an act of perceptual classification, rather than of perceptual recall. While looking at something, I classify it as red, say, or as a kestrel—and then I plug this act into the "experience operator" to form a term which refers to the experience of seeing red, or seeing a kestrel.

Now, acts of perceptual classification do not seem to differ as sharply from the experiences here referred to as do acts of perceptual recall. When I classify something as red, while looking at it, the experience this involves isn't a faint copy of the experience of seeing red. On the contrary, it is, if anything, a highlighting or intensification of that experience. The act of classification amplifies the underlying experience. In neural terms, we can usefully think of classification as occurring when some stored "template" resonates with incoming signals, and thereby reinforces or augments them.

Given this, it seems that introspective uses of phenomenal concepts will actually include the experiences they refer to, in a way that imaginative uses do not. And, to this extent, there would seem no need in the introspective case to appeal to some resemblance between exemplar and referent to fill out the quotational-indexical story.

But perhaps the introspective case presents a converse difficulty. To the extent that the act of classification intensifies the underlying experience, then it will itself be different from unintensified such experiences. It will be a vivid copy, so to speak, rather than a faint copy. So, if we want it to be an exemplar for the full range, including the unintensified experiences, there will again be some need to appeal to resemblance. On this suggestion, then, when we think "the experience: ———", and fill in the gap with an act of perceptual classification, the resulting term should be understood as referring not just to experiences of the same vivid kind as the act of perceptual classification itself, but also to any unvivid versions that resemble it.

---

20 There is a question here about whether there is any perceptual consciousness in the absence of perceptual classification. Are our sensory states conscious at all, when they aren't being "highlighted" by perceptual classification? This issue will become prominent in chapter 7. For the moment, however, I shall simply assume that we can consciously see a kestrel, or the redness of something, even when we aren't classifying it as a kestrel, or as red.
The Phenomenon of Consciousness
David Papineau

Chapter 5

The Explanatory Gap

5.1 Introduction

Joseph Levine (1983, 1993) has argued that any attempt to construct materialist reductions of phenomenal states will leave us with an "explanatory gap".

Suppose we have some theory which identifies pain, say, with some physical property, like activity in pyramidal cells. And suppose that this theory has all the empirical support it could have. As far as we can tell, pains occur when and only when pyramidal cells are active. Moreover, the two states seem to play exactly the same role in the causal scheme of things, to have exactly the same causes and effects. So, as materialists, we identify pain with pyramidal cell activity.

Even so, Levine argues, we will still lack any explanation of why pyramidal cells yield pain. There will still be a puzzle as to why it feels like that to have active pyramidal cells, rather than feeling some different way, or feeling like nothing at all. To be told that pains are always present when pyramidal cells are active is not yet to be told why those feelings should accompany those physical states.

The same point applies to materialist theories which identify pains with physically realized higher-order states, rather than with the physical realizations themselves. Suppose we were to accept, again on the basis of the fullest empirical evidence, that pain is identical with the higher-order property of having some physical state which mediates between bodily damage and avoidance behaviour. An analogous explanatory gap would still seem to remain. Why should possession of this higher-order property feel like that? We still seem to have no explanation of why the higher-order state feels that way, rather than some different way, or no way at all.

The point generalises. Take any phenomenal property C, and consider any theory that identifies it with some material property M. However well-supported this theory, it would still seem to leave us in the dark as to why M yields C. Why does it feel like that, rather than some other way, or no way at all, to have M?

Levine argues that this explanatory gap is peculiar to attempted materialist reductions of phenomenal states. Materialist reductions in other areas of science do not leave us with any similar explanatory puzzle. Once water has been identified with H2O, or temperature with mean kinetic energy, we do not continue to ask why H2O yields water, or why mean kinetic energy yields temperature. And, in general, successful materialist reductions seem to explain non-phenomenal everyday kinds in a way that removes puzzlement.

So there seems something especially baffling about the relation of phenomenal consciousness to the material world. There seems to be something about consciousness that materialist reductions cannot explain. While other everyday kinds
can be explained in material terms, consciousness seems to resist any materialist domestication.

In this chapter I shall examine this putative explanatory gap. My conclusion will be that there is nothing in it to worry materialists. I shall accept that there is indeed a kind of explanatory question which is not answered by materialist reductions of conscious properties. And I shall also accept, with some qualifications, that this marks a contrast with materialist reductions in other areas of science. But I shall show that this is just what materialists should expect. For all identity claims involving directly referring terms fail to answer explanatory questions. In as much as it is true that materialism fails to answer explanatory questions, this is simply because phenomenal concepts refer directly. Correspondingly, in as much as reductions in other areas of science do answer explanatory questions, this will be because the reduced kinds involved are referred to by description, rather than directly.

5.2 Mark Twain, Samuel Clemens and Intuitions of Gaps

An example using proper names will illustrate the central point at issue, namely, that all identity claims involving directly referring terms fail to answer explanatory questions. Consider this now well-known parable.\(^{21}\) Suppose that there are two groups of historians, one of which studies the famous American writer Mark Twain, while the other studies his less well-known contemporary, Samuel Clemens. The two groups have heard of each other, but their paths have tended not to cross. Then one year they both hold symposia at the American Historical Association, and late one night in the bar of the Chicago Sheraton the penny drops, and they realize they have both been studying the same person.

At this stage there are plenty of good questions the historians might answer. Why did this person go under two names? Moreover, why did it take us so long to realize Mark Twain and Samuel Clemens are the same person? But there is one question they won't be able to answer, because it makes no good sense: why are Mark Twain and Samuel Clemens the same person? If they are, they are, and there's an end on it.

Phenomenal concepts, like proper names, refer directly, and for this reason identity claims involving them work similarly. Let us suppose, for the sake of the argument, that we find out that pain is pyramidal cell activity. Then there are various good questions we might answer. Why do we have two different kinds of concept (phenomenal and material) for this one property? And why did it take us so long to recognize there is just one property here? But there is one question we can't answer, because it won't really make sense: why is pain the same as pyramidal cell activity? If it is, it is, and there's an end on it.

The point here is that identities need no explaining. If "two" entities are one, then the one doesn't "accompany" or "give rise to" the other—it is the other. And if it is, then there is nothing to explain. It is possible to explain why one thing "accompanies" or "gives rise to" another thing. But you can't explain why one thing is itself.

\(^{21}\) I got this story from Ned Block, and used it in when I first wrote about these matters (1993a, 1993b). He can't remember where he got it from. (Cf Block and Stalnaker, 2000, and also Block, 1978.)
Now, having made this point, I should immediately concede that I don't expect it to extinguish the underlying intuitions which fuel concern about the "explanatory gap". Let me go slowly here. I think that the Mark Twain example does provide a good model for the materialist reduction of phenomenal properties. And I therefore think that materialism leaves us in no more of an explanatory quandry than does the identification of Mark Twain with Samuel Clemens. But at the same time I recognize that this will make strike many readers as unconvincing. Isn't it obvious that materialism leaves something unexplained, in a way that Mark Twain=Samuel Clemens doesn't?

In defence of this last thought, you may want to point out that there is an obvious disanalogy between the two cases. Before the penny drops, some observant historians may become puzzled about the close proximity of Mark Twain and Samuel Clemens, and start looking for some explanation of why they always turn up in the same places. But once these historians realize there is just one person at issue, their explanatory worries will disappear. Their acceptance of the identity will quite nullify any desire for further explanation.

However, in the mind-brain case it seems quite otherwise. Even those, like myself, who are persuaded that the mind is identical to the material brain, will surely admit that they sometimes hanker for some further understanding of why brain activities should yield conscious feelings.

I concede that I sometimes find myself so hankering. But I do not think that this is because materialism is somehow explanatorily inferior to Mark Twain=Samuel Clemens. Rather, it is because materialism is so hard to accept in the first place.

The real fly in the ointment is the "intuition of distinctness" that I have mentioned in previous chapters, and which will be the focus of the next. This arises quite independently of any questions of what materialism might or might not explain. Rather, it comes from a separate source, and seduces us into thinking that phenomenal properties must be distinct from material ones. So the underlying intuition here isn't that, after we have accepted materialism, then we will be left with some worryingly unexplained business. It is an intuition that stops us accepting materialism in the first place.

Of course—and this makes it hard to keep things straight—once we have been seduced by this independent intuition of distinctness into rejecting materialism, then will indeed be faced with all kinds of unaswerable explanatory puzzles. If the phenomenal properties are distinct from material ones, then how come they always accompany each other? And how do the phenomenal properties get in on the causal act? And so on.\footnote{Recall my response to Kripke's modal argument. There too I admitted an asymmetry. Once Jane found out that Cicero=Tully, she ceased to think it possible that they could come apart; yet even convinced materialists intuitively feel that zombies are possible. However, my diagnosis there (as here) wasn't that zombies are somehow intrinsically more possible that Cicero=Tully (cf. materialism is somehow less explanatory that Mark Twain=Samuel Clemens). Rather, there is an independent source of intuition that persuades us that phenomenal properties are actually distinct from material ones, and then we infer that zombies are possible (cf. then we face an awful explanatory gap). The two cases are quite analogous.}
I am sure that it is questions like these that make people feel there is some unanswerable "explanatory gap" between brain and mind. But note how these explanatory problems presuppose that materialism is false. Correspondingly, if only we could convince ourselves properly to embrace materialism, we would be able to dismiss them as based on mistaken presuppositions. More generally, I maintain that, if we properly embraced materialism, materialism would then seem no more explanatorily puzzling than Mark Twain=Samuel Clemens.

In support of this diagnosis, it is worth pointing out that the language used to posit an "explanatory gap" often betrays an unacknowledged commitment to dualism. The problem is often posed as that of explaining how brain processes can "generate", or "cause", or "give rise to", or "yield", or "be correlated with", or "be accompanied by" conscious feelings. These phrases may seem innocuous, but they all implicitly presuppose that conscious feelings are some extra feature of reality, distinct from any material properties. And once we slip into this dualist way of thinking, then it is unsurprising that we find ourselves with unanswerable explanatory puzzles.

Given the points made in this section, the discussion in the rest of this chapter will have a slight air of unreality. In what follows, I shall consider how far different kinds of reductive these are accompanied by explanatory claims. Mind-brain reductions will be argued not to yield anything explanatory, unlike some scientific reductions, but like Mark Twain=Samuel Clemens. Further, I shall also consider whether this reflects badly on mind-brain materialism, and I shall conclude that it does not.

However, I do not think, for the reasons just given, that this has anything much to do with the vivid intuition that materialism leaves us with an "explanatory gap". This vivid feeling is a consequence of the independently motivated "intuition of distinctness", not of any explanatory deficiencies in materialism itself. To help keep this in the front of our minds, I shall avoid the term "explanatory gap" when talking about any comparative non-explanatoriness in materialism. Rather, I shall reserve this term specifically for the feeling of puzzlement we experience while we are in the grip of dualist thinking.

5.3 Reduction and Role Concepts

I said earlier that some reductions in other areas of science do yield a kind of explanation which is not provided by materialist reductions of conscious properties. It is worth considering carefully why this might be so. (In fact, by the end of this chapter this disanalogy will have been significantly qualified. But it will be expositorily helpful to start by explaining the putative basis for the disanalogy, and then adding in the qualifications as we proceed.)

At first sight, it might seem unclear how there could possibly be any such disanalogy. In both cases we start with certain pre-theoretical everyday terms, like "pain", or "thirst", in the phenomenal case, and terms like "water", "temperature", or "lightning", or so on, in other areas of science. None of these terms will have any a priori connections with terms for theoretical kinds. So in all cases the relevant pairings—of "pain" with "pyramidal cell activity", say, or "water" with "H2O"—will be established
on a posteriori grounds. The lack of any intrinsic conceptual links between any of these paired terms would thus suggest that all these reductions will present themselves as matters of brute, unexplained fact. In the scientific case, as much as with reductions of consciousness, there will be no a priori reason why some scientific and some everyday kind should go hand in hand. That will simply be how the world turns out.

But there a further circumstance which does arguably distinguish the two kinds of case. The pre-theoretical kinds involved in scientific reductions will often be picked out by "role concepts". These are concepts that pick out some property in virtue of their association with some description of a causal role. In fact there are two kinds of case here. There is the case where a concept names some first-order property, but picks it out by describing it as the first-order property which plays some specified causal role in the actual world. In chapter 3 we assumed that "water" (along with "temperature" and "lightning") work like this, naming that first-order stuff which is colourless, and so on, in the actual world. But I shall also count as a role concept any concept which names a second-order property, that is, a property of having some first-order property which plays a specified causal role. Dispositional terms, like "soluble", or "toxic", are arguably the paradigm case here.

Now, the point about such role concepts is that their association with with the relevant causal role will be an a priori matter. It is a purely conceptual matter that "water" names whichever stuff is colourless, and so on, in the actual world, or that "toxic" applies to anything that impairs bodily functions.

Suppose now that the physical facts tell us that some physical property actually fills the relevant role. It will then follow, without further ado, purely in virtue of the relevant term's a priori association with its causal role, that this physical property is, or realizes, the referent of the relevant term. Once we are shown that H2O is colourless, and so on, it follows a priori that H2O is water. Once we are shown that carbon monoxide impairs bodily functions, we know a priori that CO realizes toxicity.

In fact, the supposition just made here, that the physical facts can tell us that some physical property actually fills some role, is a significant extra assumption, over and above the assumption that our pre-theoretical kind is picked out by some role concept. For it should not be taken for granted that all causal roles mediate between physical inputs and outputs (which is what we need if the physical facts alone are to tell us that some physical property actually fills the relevant role). Still, I propose to let this worry pass for the time being. I shall return to it in section 5.6 below.

Even if we make this supposition, that the physical facts alone imply a priori that water is H2O, and that CO is toxic, and so on, this doesn't of course mean that our overall reduction becomes entirely a priori. It is an empirical matter, which certainly does not follow from the definitions of "water" and "toxicity" alone, that H2O is colourless, odourless and tasteless, or that CO impairs bodily functions. The idea is rather that, once we have established these physical facts, then nothing more is needed, beyond conceptual analysis, to reach the reductive claims. In this sense, the reductive claims follow a priori from the physical facts alone.

Given our assumptions, it will then also follows there is a sense in which physical reductions will explain such things as why H2O is water, or why CO realizes toxicity.
For the full physical story will presumably show us how it is that H2O is colourless, odourless, and tasteless, and how it is that CO impairs bodily functions. And in this sense the physical story will explain the connection between the scientific kind and the everyday kind. More generally, whenever a physical reduction tells us that some physical property is identical with, or realizes, some everyday type, there would seem to be room for an explanation of why the physical property in question satisfies the descriptions associated a priori with the relevant everyday concept.

Now, none of these points about scientific reductions apply to identifications of phenomenal concepts with material properties. This is because phenomenal concepts refer directly, and not via any a priori association with causal roles. So, however many physical facts we are given, there is no question of inferring a priori that some phenomenal property is identical with, or realized by, some physical property. The physical story may tell us that certain physical properties play certain causal roles. But this won't imply anything a priori about the relation between those physical properties and any phenomenal properties, given that there are no a priori links between phenomenal concepts and causal roles.

Nor, by the same coin, will there be any question of explaining why some physical property, like pyramidal cell activity, is identical to, or realizes, some phenomenal property, such as pain, along the lines of the above explanation of why why H2O is water, or CO realizes toxicity. These explanations of how physical states reduce everyday kinds hinged on the association of those kinds with causal roles. Insofar as there are no causal roles associated a priori with phenomenal concepts, no physical story, however full, is going to explain in this way why certain physical activities yield conscious states.

5.4 Identities and Descriptions

This might seem like an embarrassing admission from a materialist about consciousness. If we can't explain why the brain generates feelings, in the way that we can explain why H2O is water, or why CO realizes toxicity, then doesn't it follow that consciousness will always remain a mystery, by comparison with other scientific topics?

No. Even if there is this difference between mind-brain identities and other scientific reductions, this does nothing to discredit mind-brain identities. Claims of mind-brain identity do not call for explanation. Recall my analogy with Mark Twain. I say that once you really accept that pain, say, really is some material M, then you will see that this requires no more explanation than does Mark TwainSamuel Clemens. Identities need no explaining.

This claim that identities need no explaining may seem to be belied by the scientific examples discussed a moment ago. Didn't I just admit that physics can yield an explanation of why H2O is water, of why mean kinetic energy is temperature, and so on? If we can explain these identities, then why shouldn't we be able to explain mind-brain identities?
But these are not really explanations of identities. We aren't explaining why H_2O is water—that is, why it is the stuff—H_2O—which happens to play the role of being colourless, odourless, and so on. This would be to try to explain why H_2O is itself, which would be misplaced. Rather, we are explaining why H_2O is colourless, odourless and tasteless. We are explaining why H_2O satisfies the descriptions associated with the other way of picking it out, as "water". This is a perfectly good thing to explain, and I allowed above that physics can explain such things. But this is not a matter of explaining an identity—of explaining why some entity is itself—but rather of explaining why some entity possesses certain further attributes.

Consider this analogy. I can explain why Tony Blair is the British Prime Minister. Thus, I might point out that the Labour Party won the last election, and that Blair was selected as leader of the Labour Party because many of the Parliamentary Labour Party thought he was the candidate most likely to win the election, and . . . (where the exact answer I give will depend, as with all explanations, on the interests and epistemic needs of my audience).

However, I am clearly not here explaining why Tony Blair is himself. Rather I am explaining why he has a certain attribute, that of being the British Prime Minister. We should not be misled by the surface form of the explanandum—Tony Blair is the British Prime Minister. Even if this explanandum does imply a personal identity, it is not this identity we are explaining, but rather the attribution which is also implied, that of British-Prime-Ministerhood to Tony Blair.

Similarly with explanations of H_2O=water and similar scientific identities. We are not in fact explaining a property identity, but rather the implied attributions of colourless, odourlessness, and tasteless to H_2O.

This is why there is nothing analogous to explain in the mind-brain case. Phenomenal properties refer directly, and not by invoking any further features of their referents. So identity claims involving them will in a sense be "brute" identity claims. They will simply say that one thing—a phenomenal property—is identical with another—a material property—without alluding to any further features the phenomenal property may possess. The phenomenal concept will simply name the phenomenal property, so to speak, without describing it further. And since brute identities state no other fact, except to say of something that it is itself, they raise no issues of explanation.

5.5 Does Physicalism Require the Physical Truths to Imply all the Truths?

Notwithstanding the points made so far, some philosophers hold that materialism must offer something more than brute mind-brain identities. They argue that materialism is committed to the view that that all truths—including all truths about the mind—must follow a priori from the physical facts. (Cf. Jackson, 1998.) If this is right, then materialists cannot rest with unexplainable mind-brain identities. They have an

---

23 More precisely, this is normally argued to be a commitment of physicalism, rather than materialism. Still, they come to the same thing, given that my materialism takes all first-order properties to be physical.
obligation to portray the phenomenal realm in a way that will make it a priori comprehensible how the physical world can yield phenomenal minds.

This line of thought, however, need not detain us long. For materialists, at least those of inflationary stripe, will simply reject the above characterization of materialism. That is, they will simply deny that their materialism requires all truths to follow a priori from the physical truths.

Materialism would only require this if all non-physical concepts picked out their referents via descriptions of causal roles. On this conception of materialism, all concepts of properties would either be overtly physical—that is, pick out their referents as physical properties—or specify causal roles which mediate between physical inputs and outputs. Since materialism does plausibly require that all first-order properties are physical, it would then follow, with equal plausibility, that the full physical story will a priori fix the complete inventory of satisfiers of both physical and non-physical concepts, that is, all truths, however formulated.

However, the claim that all non-physical concepts refer via association with causal roles is precisely what inflationary materialism denies. Inflationary materialists take phenomenal concepts to refer directly, in their own right, and not via any specification of such roles. So inflationary materialists will see no reason to accept, even given their materialism, that the satisfaction of such concepts can be inferred a priori from any physical story, however full. Thus, as materialists, they will take the phenomenal concept of pain, for example, to refer to some material state. But since they also hold that this phenomenal concept has no a priori connections with causal roles of any kind, so there is no question of inferring its satisfaction a priori from any physical story.

In discussing Kripke's modal argument in chapter 3, I had occasion to criticise the idea that, whenever two directly referring terms refer to the same thing, it must be a priori knowable that they do so. This same dubious assumption seems to me to lie behind the contention that materialism requires all truths to follow a priori from the physical truths. For this contention simply rules out the possibility that all first-order properties may be physical, thus ensuring materialism, and yet two a priori distinct concepts may both refer directly to such a property. I still see no reason to accept the thesis that a priori distinct concepts cannot both refer directly to the same thing. As before, this thesis seems to me to depend on archaic assumptions about unmediated mental acquaintance, assumptions which have been amply discredited by recent thinking about reference.24

24 Ned Block and Robert Stalnaker (2000) point out that something else is odd about the view that the physical truths imply all the truths. Even if all concepts were role concepts, would all truths really follow? What about the possibility that some role is filled twice over, by a non-material property as well as a physical one, as on the overdeterminationist "belt and braces" view discussed in chapter 1? Then there would be some truths, namely, about the non-material realizers of the role, which couldn't be derived from the physical truths and conceptual analysis. You might want to reply that such non-material fillers would themselves be inconsistent with materialism. (After all, remember that the contentious issue is not whether all truths follow from the physical truths, but whether materialism is committed to this.) Still, it is not clear that this serves. We can agree that materialism implies that there are no non-material properties. But does materialism imply that physics implies this? It seems not. Physics itself does not plausibly say anything about the non-existence of non-material stuff, whether or not it is conjoined with materialism.
5.6 A Epistemological Gap

Let me now consider one last worry that might be associated with the non-explanatoriness of materialist reductions. This is that the non-explanatoriness may remove our reason for believing in materialist reductions in the first place. Perhaps, this objection can concede, there wouldn't be anything wrong with materialist reductions of phenomenal states, if only we had reason to believe them. But how can we be in a position to believe them, if they requires us to believe brute and unexplained identities which cannot be derived a priori from the physical facts?

The thought here would be that any epistemological access to a materialist reduction must proceed via by showing how some physical property satisfies a causal role. (Levine, 1993.) The causal role will be associated a priori with the reduced kind, and then the physical facts will then show us how some physical state realizes that role. This, so the argument goes, is how we find out that H₂O is water, or that CO realizes toxicity. Yet inflationary materialists admit that we can't similarly find out that some physical property realizes the role associated a priori with some phenomenal concept, since they maintain there are no such roles associated with phenomenal concepts.

A first materialist response to this challenge would be that this kind of role-realizer discovery isn't the only possible epistemological basis for believing in a reduction. We can also have direct evidence for a reductive conclusion. Consider the straightforward causal argument for mind-brain identities adduced in chapter 1. This owed nothing to any a priori analyses of the reduced phenomenal kinds as involving a priori roles. Indeed, for that matter, consider personal identities like Mark Twain=Samuel Clemens. These can't be epistemologically based on the uncovering of role-realizer structures, since there are no roles in play here. Yet our knowledge of them seems none the worse for that.25

Still, the objector might persist, knowledge of role-realizer structures is epistemologically crucial for serious scientific reductions. Maybe familiar everyday identities involving spatio-temporal particulars, like Mark Twain=Samuel Clemens, can be happily accepted on brute correlational evidence alone. But when it comes to the identifications of natural kinds with unfamiliar theoretical kinds, surely we need something more substantial, something that will enable us to understand why the theoretical kind realizes the everyday kind. And this will require role-realizer structure, of the kind unavailable in the mind-brain case. This is why we find claims of mind-brain identity so unconvincing, by comparison with established scientific reductions. Or so this objection goes.

At this stage a different line of response is open to materialists. They can query whether established scientific reductions all conform to the role-realizer model in the first place. (Cf. Block and Stalnaker, 2000.) So far I have not disputed this contention. But a moment's thought will show it is highly dubious. Take water=H₂O. This has been known since the middle of the nineteenth century. But there was no possibility of any physical explanation of why H₂O is colourless, or has other water-identifying

25 Sturgeon (1994) argues, against this, that all concepts other than phenomenal concepts allow the uncovering of role-realizer structures.
properties, until well into this century, with the advent of quantum mechanics. (Indeed I doubt that anybody has even now bothered to carry out any such explanations in any detail.) So the recognition that water is H₂O could not possibly have depended on any physical explanation of how H₂O realizes some a priori water role. Instead it must have been based on more direct evidence.

Similarly, I would suggest, with other scientific reductions. The recognition that lightning is atmospheric electrical discharge derived from experiments, like Franklin's, which simply showed that electrical discharges occur when lightning does. It owed nothing to any physical explanation of how electrical discharges produce the effects associated with lightning. Again, the recognition that certain substances, like CO, realize toxicity, will not (thank goodness) standardly wait on a full physical explanation of how they produce their noxious effects, but will simply be accepted on more direct grounds.

A point flagged earlier is relevant here. In section 5.3 I pointed out that, even if our pre-theoretical concepts of water, lightning, temperature, toxicity, and so on, are role concepts, it doesn't follow that the relevant roles will be specified entirely in terms of physical inputs and outputs. And, on reflection, it is clear that any roles associated with these concepts will not be so purely physically specified. "Colourless", "odourless", "flashing across the sky", "causing heat sensations", and "impairing bodily functions" are not concepts used in physics. They are phenomenal or perceptual concepts, not concepts which appear in the vocabulary of physical science.

The assumption that all the relevant roles must be purely physically specified is presumably an upshot of the view of reference mentioned in the last section. This view—the one that the anti-materialists tried to foist on materialists—had it that all non-physical concepts must pick out their referents via purely physically specified causal roles. Still, as before, materialists will simply reject this view. I have already emphasized the possibility of referring to material states using phenomenal concepts. Also relevant at this stage is the possibility of referring to material states using perceptual concepts (such as "flashing across the sky", or "has stopped breathing"). All phenomenal concepts, and most perceptual concepts, play no role in physics itself, yet will be regarded by materialists as perfectly good ways of referring to real material properties, and so as perfectly good ways of specifying the inputs and outputs of causal roles.

This means that, even if some pre-theoretical kind is picked out by a role concept, the relevant role is likely to mediate between phenomenally and perceptually specified inputs and outputs, rather than physically specified ones. And this further means that, even if some kind is picked out by such a role, we won't be able to identify a realizer a priori solely on the basis of physical information. If the inputs and outputs aren't physically specified, physical information alone cannot tell us a priori how they are realized. Rather, at some point in accepting the reduction, we will have to embrace some brute phenomenal-physical identities, or brute perceptual-physical identities, purely on the basis of direct correlational evidence.

---

26 Note that I am here making use of the assumption, defended in 4.7 above, that perceptual concepts typically refer directly, and not because of a priori associations with roles.
I do not necessarily want to maintain that role-realizer structure plays no part in establishing orthodox scientific identities. I do think that our pretheoretical concepts of "water", "lightning", "toxicity", and so on, are role concepts: they pick out their referents a priori as entities with certain perceptual and phenomenal causes and effects. Moreover, to the extent that this is so, there will be something explanatory about claims that such pretheoretical kinds are realized by physical types: these reductive claims will have the potential to explain why those physical types go with the physical correlates of those perceptual and phenomenal causes and effects.

Still, insofar as these role-specifying perceptual and phenomenal types are not themselves physical types, physics alone will not be able identify the relevant physical correlates a priori. At some point the epistemological buck will have to stop. The physics and the conceptual analysis will fail to meet, and we will have to use direct correlational evidence to equate physical states with phenomenal or perceptual ones.

The overall suggestion at issue in this section is that mind-brain identities must be epistemologically underprivileged, by comparison with scientific reductions, since they cannot be derived a priori from physics. However, we can now see that there is no qualitative epistemological difference of this kind between mind-brain identities and reductions in other areas of science. Maybe scientific reductions make some use of role thinking. But in the end they rest on the acceptance of brute identities, just like mind-brain reductions.

I take it that no-one will want to use this point to cast epistemological aspersions on standard scientific reductions. If we know anything, we know that water=H₂O. So, by the same coin, we have no reason for epistemological despondency about mind-brain reductions. They may require us to embrace brute identities, but so does the rest of science.

5.7 Conclusion

In this chapter I have conceded that mind-brain reductions are non-explanatory. Conceptual analysis plus physical facts cannot show us why certain brain states constitute certain feelings. This is because phenomenal concepts are not associated with roles, and so there is no question of showing how certain entities fill those roles.

Still, I have argued that this does nothing to discredit mind-brain identities. Other familiar identities are equally non-explanatory. For example, brute personal identites, like Mark Twain=Samuel Clemens, convey no explanatory information. Maybe there is something genuinely explanatory about scientific reductions, like water=H₂O. But even these, as we have just seen, will at some point rest on brute non-explanatory identifications of non-physical with physical kinds.

Some readers may feel that something must have gone wrong here. For surely, if we stop to think about it, and get away from the technicalities, there is something distinctively problematic about mind-brain identities. To return to the intitial concern about the "explanatory gap", won't it always remain mysterious how brains give rise to pains, and colour experience, and all the rest of the rich phenomenal life we so enjoy?
How could squishy grey matter possibly do all that? Everybody surely feels this puzzle. Yet we feel no corresponding mystery about Mark Twain=Samuel Clemens.

I agree that there is something distinctive about the mind-brain case. But this doesn't show that there is anything wrong with the analysis in this chapter. Rather, it shows that something else makes us puzzled about mind-brain identities. As far as explanatoriness goes, mind-brain identities are no worse than many other respectable identities. If explanation were all that mattered, we wouldn't find the mind-brain relation any more mysterious than Mark Twain=Samuel Clemens.

However, there is something else that matters, namely the independent "intuition of distinctness" that I have mentioned before. This doesn't apply to Mark Twain=Samuel Clemens, which is why we have no difficulty with this identity. But it does apply to the mind-brain relation, and this is why we feel it is different.
6.1 Introduction

Let me now focus on the intuition of distinctness itself. In my view, this is what makes the mind-body problem seem so intractable. Even given all the arguments, intuition continues to object to mind-brain identity. 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?

In this chapter, I shall try to explain this intuitive resistance to materialism about the mind. I think there is indeed something special about the mind-brain relation. It generates this overwhelming intuition of distinctness. Even convinced materialists are likely to feel the pull of this intuition. I know that in my own case it continues to press, despite any amount of immersion in the arguments of the previous chapters.

I don't think that this contrary intuition discredits materialism, because I think it is mistaken. But I do think that it needs to be recognized and explained. Materialism will remain unconvincing until this anti-materialist intuition is laid to rest. An intuition on its own does not amount to an argument. But, still, it is a striking feature of the mind-brain relation that it does generate this contrary intuition, and a full understanding of the subject ought to explain why this is so. Even if materialism isn't mistaken, its defenders at least owe an explanation of why it should seem so mistaken.

6.2 Is an Explanation Already to Hand?

Despite some contrary suggestions I have made in previous sections, some readers may feel that ample materials for such an explanation are already to hand. Consider the anti-materialist arguments discussed in previous chapters. I have argued that they do not establish their anti-materialist conclusions. Even so, might they not still be responsible for the impression that materialism is false? Even if the anti-materialist arguments are unsound, they aren't obviously unsound. So perhaps the thinking behind these arguments is itself responsible for the widespread credence given to denials of mind-brain identity.

Thus we had Jackson's knowledge argument, Kripke's modal argument, and the argument from the non-explanatoriness of mind-brain identities. These arguments may be flawed, but they can all seem highly convincing at first sight. So why not simply explain any intuitions of mind-brain distinctness as upshots of the persuasiveness of these arguments?
I want to resist this suggestion. I don't think any of these arguments get to the heart of the crucial intuition. In the next section I shall back up this claim by reminding readers how these arguments apply equally well to cases where we do not find any intuition of distinctness. So the intuition of phenomenal distinctness must have some other source.

Of course, given that there is this intuition of distinctness, it cannot help but lend apparent weight to the anti-materialist arguments. For the intuition will support the conclusions of these arguments, even if it stems from a quite different source. It will make the anti-materialist arguments seem more convincing than they deserve to seem, simply because it portrays them as leading to the truth.

But this is different from saying that the arguments explain the intuition. Indeed, it says precisely the opposite, since it is only because the intuition of distinctness has a different source that it can serve to bolster the arguments. So there must be some other explanation for the compelling intuition that the phenomenal mind must be something extra to the brain.

6.3 The Independence of the Intuition

Let us first check that the intuition of distinctness really does have an independent source, apart from the plausibility of the earlier anti-materialist arguments. My strategy, as I said, will be to remind you that analogues of these arguments apply equally well to cases where we do not find any corresponding intuition of distinctness.

Let me start with the argument from non-explanatoriness. I take the last chapter already to have established that this constitutes no good objection to materialism. But that is not the present issue. Rather, the current question is whether this non-explanatoriness can account for the compelling (even if mistaken) intuition that mind must be distinct from brain.

Here I need only repeat the point with which I finished the last chapter. Materialism is indeed in one respect non-explanatory. Since phenomenal concepts do not allude to causal roles, there is no question of explaining how certain physical states play those roles. But other familiar identities are quite comparable in this respect. Many identities involving ordinary names are "brute identities", free of any allusions to causal roles. Yet we feel no persistent intuition of distinctness in these cases. So the mind-brain intuition of distinctness must depend on something else.

Now consider Kripke's modal argument. This started from the apparent contingency of the relation between phenomenal mind and brain; it continued by pointing out that this can't be explained in terms of some genuine possibility of the relevant brain state not satisfying some phenomenal role concept; and it concluded that phenomenal mind and brain must therefore be genuinely distinct. As before, the unsoundness of this argument is not at issue: we have already seen how other respectable identities involving non-role concepts will also yield brute illusions of contingency. The issue is rather whether the brute appearance of contingency can account for the illusion of mind-brain distinctness.
To see that it cannot, we need only consider personal and scientific identities again. Recall the earlier example of Jane, who had picked up the names "Cicero" and "Tully". Prior to Jane's discovery that these are two names for one person, she could well think that Cicero might not be (or indeed is not) Tully, even though there is no genuine possibility in the offing which might be viewed as giving content to her thought. Yet, when she does discover that Cicero=Tully, she will have no residual intuition that Cicero can't really be Tully, akin to the intuition of distinctness found in the mind-brain case. So again this latter intuition must derive from something more than that the ability of direct names like "Cicero" and "Tully" to generate appearances of contingency.

Let me now turn to Jackson's argument. When discussing this earlier, in chapter 2, I made no explicit mention of the distinctive intuition of mind-brain distinctness. Still, some readers may feel that it is precisely Jackson's points that hold the key to this intuition. For, even if Jackson's argument fails to discredit materialism, it did establish that we have two very different ways of thinking about phenomenal/material properties. Perhaps this extreme difference at the conceptual level is what we need to account for the intuition of distinctness.

The thought here would be that this persistent intuition arises simply because our two ways of thinking about material properties are so very different. Presenting a feeling in phenomenal terms, as a feeling, and presenting it in material terms, as a material state, are qualitatively quite different mental acts. Maybe this striking difference in the two modes of reference is the reason we find it so hard to believe that they pick out the same property.

In this connection, it is arguable that the proper name and scientific cases do not display the same kind of conceptual dualism. While these may involve two personal names ("Cicero", "Tully"), or two kind terms ("water", "H2O"), the names in these pairs will be of the same general type. Given this, so the argument would go, we have no difficulty accepting that "Cicero" and "Tully", or "water" and "H2O", can name the same things. It is easy enough to grasp the idea that two terms from the same general stable can name the same entity. By contrast, it may not be so easy to stomach an identity involving two radically different terms, one phenomenal and one material.

Now, I think this suggestion is close to the truth. Later in this chapter I shall argue that the persistent intuition of mind-brain distinctness is indeed due to peculiarities of the dualistic conceptual structure by which we refer to phenomenal/material properties. But I do not think that it is enough just to point to the existence of this dualism. It is not just that we have two different ways of thinking about phenomenal/material properties. This alone does not explain the intuition of distinctness. Rather, the intuition derives from a special further feature of our dualistic conceptual structure. Or so I shall shortly argue.

But first I would like to refer back to chapter 4 in support of my claim that conceptual dualism in itself is not enough to account for the intuition of distinctness. Remember my discussion of perceptual concepts. These were concepts associated with perceptual classification and perceptual recall, and referred to features of the external world, like birds, or elephants, or colours of objects, and so on.
Now, such perceptual concepts are themselves radically different from other ways of referring to things in the external world. To think of a kestrel visually is a quite different mental act from thinking about it in some way that would be available to someone blind from birth. Indeed, as I pointed out in chapter 4, the main features of Jackson's argument apply to perceptual concepts as much as to phenomenal concepts. In general, you can only have perceptual concepts for those simple things that you have perceived previously. No amount of book learnin' will tell you how to think of something visually, if you have never visually perceived anything of that kind before.

Yet, despite this, I take it that we don't find a intuition of distinctness about the referents of perceptual and other concepts, corresponding to intuitions of phenomenal-material distinctness. Suppose John has a visual concept of a kestrel, of the kind unavailable to a congenitally blind person, yet can also think about this bird non-visually, in a way that would be available to a blind person. And suppose that John yet doesn't know that these two concepts refer to the same bird. Then he discovers that they do: his visual kestrel—the bird like that—is the same as his non-visual kestrel. I take it that he won't at this stage feel any residual intuition of distinctness. He won't feel that there is something mysterious about the non-visual bird "giving rise" to the visual one. Yet this identity involves concepts of radically different kinds, just as much as any phenomenal-material identities. So the mystery peculiar to the mind-brain relation must derive from something more that conceptual difference per se.

6.4 Nagel's Footnote

In footnote 11 of "What Is It Like to Be a Bat?" (1974), Thomas Nagel considers Kripke's challenge to mind-brain identity, and suggests that it might be met by distinguishing between "perceptual" and "sympathetic" imagination. As he puts it:

"To imagine something perceptually, we put ourselves into a conscious state resembling the state we would be in if we perceived it. To imagine something sympathetically, we put ourselves in a conscious state resembling the thing itself. (This method can only be used to imagine mental events and states—our own or another's.) When we try to imagine a mental state occurring without its associated brain state, we first sympathetically imagine the occurrence of the mental state; that is, we put ourselves in a state that resembles it mentally. At the same time, we attempt to perceptually imagine the non-occurrence of the associated physical state, by putting ourselves into another state unconnected with the first: one resembling that which we would be in if we perceived the non-occurrence of the physical state. Where the imagination of physical features is perceptual and the imagination of mental features is sympathetic, it appears to us that we can imagine any experience occurring without its associated brain state, and vice versa. The relation between them will appear contingent, even if it is necessary, because of the independence of the disparate types of imagination." (My italics.)

Now, some of this repeats thoughts I have already dismissed as insufficient to explain the intuition of distinctness. But the phrases I have italicised point to a rather different explanation of this intuition.
In the first instance, Nagel can be read as simply pointing to the possibility of inflationary materialism. True, I have explained inflationary materialism in terms of a contrast between "phenomenal" and "material" thinking, rather than between "sympathetic" and "perceptual" thinking. Still, if we equate his "sympathetic" with my "phenomenal", and now include perceptual concepts as a special case of material concepts (cf. footnote 2 in chapter 2 above), we can see him as advocating a special case of inflationism: materialists should appeal to the fact that we can think about conscious properties in two different ways—one phenomenal and one not.

Nagel then points out that this in itself enables materialists to explain why it seems possible for mind and brain to come apart, consistently with their commitment to this not being possible. The two distinct concepts of conscious feeling can flank a term for non-identity, and there we are. Still, as I have argued, this in itself doesn't explain why intuition so continues to resist mind-brain identities, even in the face of strong arguments. Exactly the same explanation of illusions of contingency is available whenever we have two terms for one thing ("Cicero" and "Tully"), yet in these other cases we are perfectly ready to accept the identity once we are shown the evidence.

Again, Nagel makes the point that these two ways of thinking about phenomenal/material properties involve two markedly different mental powers—sympathetic imagination on the one hand, and perceptual imagination on the other. This corresponds to the thought considered at the end of the last section, that conceptual dualism per se might explain the intuition of mind-body distinctness. But this thought too was found wanting, given that other kinds of conceptual disparity seem not to generate any intuition of distinctness.

However, Nagel also says something further. He observes that when we "imagine something sympathetically, we put ourselves in a conscious state resembling the thing itself. (This method can only be used to imagine mental events and states—our own or another's.)"

Though Nagel does not develop it, this seems to me the crucial point. Uses of phenomenal concepts resemble the conscious properties being referred to. Moreover, this kind of resemblance between concept and object is peculiar to uses of phenomenal concepts. So far, we have not appealed to this in trying to understand our peculiar attitude to mind-brain identities. I think that it holds the key to the intuition of distinctness.

6.5 The Antipathetic Fallacy

I have argued that the important intuition of distinctness is nothing to do with modality or explanation. It is simply the intuition that, in the actual world, phenomenal states are different from material states. Of course, if you accept this intuition, then (a) you will think, modally, that phenomenal and material states could be decorrelated from each other, and then (b) you will hanker for an explanation of why they should be found together in the actual world. But I maintain that these modal and explanatory

---

27 This particular anti-Kripkean suggestion of Nagel's has been developed at length by Christopher Hill (1997). See also Hill and McLaughlin, forthcoming.
concerns are consequences of the prior thought that phenomenal and modal properties are actually different, not its cause.

So why do we have this intuition of actual difference, if it is independent of modal and explanatory issues? I think it derives from the special feature of phenomenal concepts to which Nagel draws our attention, that their uses resemble the conscious properties being referred to.

Consider the two ways in which phenomenal concepts can be deployed, that is, in imaginative recreations and in introspective classification. Both these exercises of phenomenal concepts have the unusual feature that we use versions of 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. And, when we classify an experience, we amplify it into a "vivid copy" of itself.

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 imagine 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 the exercise of some material concept which might refer to just the same phenomenal material state. No similar feelings there. To think of pyramidal cells being active, or of some-physical-state-which-arises-from-damage-and-causes-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 an intuitive sense in which exercises of material 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 material concepts, we are not thinking about the experiences themselves. After all, don't the material modes of thought "leave out" the experiences, in a way that our phenomenal concepts do not? And doesn't this show that the material 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 what lies behind the inescapable conviction that the mind must be extra to the brain. (Consider again how the standard rhetorical ploy juxtaposes phenomenal and material concepts: "How could this panoply of feeling arise from mere neuronal activity?")

However, this line of thought is a fallacy (indeed a species of use-mention fallacy, which elsewhere I have dubbed the "antipathetic fallacy"—Papineau, 1993a, 1993b). There is a sense in which material 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 material concepts "leave out" the feelings in the sense of failing to mention 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 mention, phenomenal concepts are very much the exception. So we shouldn't conclude on this account that material concepts, which work in the normal way of most concepts, in not using the states they mention, fail to refer to those states.

This then offers a natural account of the intuitive feeling that conscious experiences must be distinct from any material states. This feeling 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 bear a phenomenal resemblance. And this then creates the fallacious impression that other, material ways of thinking about those experiences fail to refer to the felt experiences themselves.  

6.6 Do Phenomenal Concepts Resemble their Objects?

The diagnosis I have offered appeals to Nagel's observation that "when we imagine something sympathetically, we put ourselves in a conscious state resembling the thing itself" (my italics). Uses of perceptual concepts resemble the conscious feelings they refer to, and this is why other concepts of conscious states can seem to "leave out" the feeling themselves.

However, there is room to doubt Nagel's original contention. Do exercises of sympathetic imagination really resemble the experiences imagined? When I imagine a pain, for example, there is indeed something conscious going on. But surely this conscious act does not feel the same as a real pain. It doesn't hurt, or make me desire its cessation. So why say it resembles the real pain?

The point generalises. Even it is like something sympathetically to imagine experiences, these acts of sympathetic imagination are surely phenomenally quite different from the experiences themselves. Nobody is likely to muddle up an imaginative act with an actual experience. Far from resembling each other, they seem quite different in kind.

This is a reasonable challenge, which raises a number of interesting issues. Even so, I do not think it discredits my diagnosis of the antipathetic fallacy. Let me make three connected points in support of this claim.

First, even if imaginative uses of phenomenal concepts do not resemble the experiences imagined, these are not the only uses of phenomenal concepts. We can also think about experiences phenomenally by using introspective classification. I classify some current experience as a pain, or as seeing red, or as seeing a kestrel, and am thereby able to think about that experience.

---

28 Brian Loar also points to this possible explanation of anti-physicalist intuitions in his "Phenomenal States" (1990, p. 90).
Now, I take it to be uncontentious that these uses of phenomenal concepts resemble the experiences they refer to. As I put it back in chapter 4, these introspective uses actually include the experiences themselves, while simultaneously highlighting or intensifying them. They are vivid copies of the experiences, rather than faint ones. So there is no doubt that the act of introspectively classifying a pain, say, will resemble a pain. Given that this act includes the pain, it will feel like a pain. It will hurt, and make me want it to go away.

So perhaps I should have restricted my diagnosis of the antipathetic fallacy to introspective uses of phenomenal concepts, and ignored imaginative uses. That is, I could have said that the impression that material thinking always "leaves out" the experience itself arises specifically when we compare such material thinking with introspective thinking. Since there is no doubt that introspective references feel like the experiences referred to—after all, they include them—there is no room here to query the initial idea that phenomenal thoughts about experiences resemble the experiences themselves.

To this extent, sympathetic imagination is not the best case for my the diagnosis of the antipathetic fallacy. Nagel's emphasis on imagination thus points us in somewhat the wrong direction. A far more obvious source for antipathetic confusion is the resemblance of conscious states to introspective thoughts, rather than to imaginative thoughts.

Still, having made this concession, I do think I can after all defend the idea that imaginative thoughts, as well as introspective thoughts, resemble their conscious objects, and that this also plays a part in seducing people into the antipathetic fallacy.

This brings me to my second point—namely, that the issue which matters here is whether normal people judge there to be a resemblance between imaginative acts and conscious feelings, not whether these judgements are defensible. For it will be individual judgements of resemblance that push people towards the antipathetic fallacy, whether or not these judgements pass any further tests of philosophical respectability. As soon as anyone feels that imaginative acts resemble their conscious objects, then this in itself can start pushing them towards the fallacious conclusion that material thoughts, by contrast, "leave out" the conscious states. (As will similar reflections about introspective acts—I shall now take it for granted that reflection on introspection, at least, does fuel antipathetic confusion.)

I contend that people do make such subjective judgements about imagination. That is, I take it to be a common everyday thought that imaginative acts resemble the experiences imagined, even if it is possible to raise philosophical queries about such resemblances. Perhaps visual imagination provides the clearest examples. Imagining seeing a red square resembles actually seeing a red square. Visually imagining isn't exactly like seeing, of course, but there is an obvious sense in which imagining and seeing are phenomenally similar from the subject's point of view. Nor is the phenomenon restricted to the visual realm. Even if imaginings of pains don't really hurt, they can share some of the phenomenal unpleasantness of real pains. An

29 Not that there is any definite indication that Nagel was trying to identify my antipathetic fallacy in his footnote 11. On the other hand, there is nothing else in his footnote that hinges on issues of resemblance between conscious states and thoughts about them.
imagined pain may not be unpleasant in just the same way as a real one, but it can still make you feel queasy, or make you twitch, or make the hairs in your neck stand on end. Again, imagining tasting chocolate feels akin to actually tasting chocolate. Even if it's not as nice, it can still make your mouth water.

So I contend that, in cases like these, it is very natural for people to think that imaginative acts phenomenally resemble the experiences imagined. And then, I say, these thoughts will help push them towards the antipathetic mistake that material thoughts "leave out" experiences.

Indeed—and this is now my third and final comment—these subjective judgements of phenomenal resemblance are arguably built into the very structure of our phenomenal concepts themselves.

Recall the final account of phenomenal concepts that I defended back in chapter 4. On that account, an imaginative or classificatory act gets plugged in to the "experience operator", and the resulting term is then understood as referring to that conscious experience which resembles this imaginative or classificatory act. If anything like this is right, then the idea that uses of phenomenal concepts resemble their conscious referents will not be some stray reflection that might or might not strike somebody musing on the nature of consciousness. Rather, an awareness of such resemblances will be a precondition of using phenomenal concepts in the first place. Competence in phenomenal thinking will depend on your sensitivity to resemblances between phenomenal thoughts and their objects.

This would help explain the pervasiveness of the antipathetic fallacy. Anybody who is able to think of phenomenal states as phenomenal will therewith think of these thoughts as resembling their objects, in a way that material thoughts cannot. They will thus be ripe for the fallacious conclusion that the material thoughts "leave out" the experiences, and will become intuitively convinced that conscious experiences cannot be identical with any material states.
The Phenomenon of Consciousness
David Papineau

Bibliography


Block, N. Forthcoming. "What is the Real Epistemic Problem about the Physical Explanation of Consciousness?"


Hill, C. and McLaughlin, B. Forthcoming. "There are Fewer Things in Reality than are Dreamt of in Chalmers's Philosophy". Philosophy and Phenomenological Research.


Nagel, T. 1974. "What is it Like to be a Bat?" Philosophical Review 83 435-450.


