we are conscious because our brain have qualia.

12

Phenomenal consciousness

Everything that you experience, Not conscience.

qualia, physical properties?
According to physicalism: yes, everything is physical

> How can technicolor phenomenology arise from soggy grey matter?
>
> —Colin McGinn

Broadly speaking, phenomenal consciousness challenges physicalism in two ways. First, it presents a **metaphysical challenge**: Can qualia be accounted for in purely physical terms? Second, it presents an **epistemological** (认识论) **challenge**: even if we accept that qualia are in fact physical, can we understand how the brain generates phenomenally conscious experiences? As far as I know, the first person explicitly to distinguish the metaphysical questions about qualia from the epistemological ones was Joseph Levine (1993). The distinction is, however, implicit in some of Frank Jackson's earlier remarks (1982).

One of the most powerful ways to press the metaphysical challenge to physicalism is Jackson's **knowledge argument** (see especially Jackson 1982; 1986). I present the knowledge argument in Section 12.1, and discuss a series of replies to it in Section 12.2. The epistemological challenge is introduced in Section 12.3. Following standard usage, I call that challenge the *explanatory gap argument*. (The name is slightly misleading as there are in fact a number of arguments at issue.) In Section 12.4 I discuss a variety of replies to the explanatory gap argument, concluding that there really is a serious problem here. In Section 12.5 I briefly discuss functionalism and phenomenal consciousness. Functionalism deserves a section of its own because of the very significant role it has played in recent philosophy of mind. I then sum up my own views in a brief conclusion.

12.1 The knowledge argument

Mary is a super-smart scientist who has been raised from birth in a black and white room. In the room she learns *all* the physical facts of relevance to human color vision—she learns all about the physics of light, the optical properties of the eye, and the anatomy and physiology of the visual system. One day she is released from her black and white prison and sees a ripe tomato in good light. 'Wow!', she says, 'Now I know what red looks like.'

When Mary was in the black and white room she knew all the physical facts of relevance to human color vision. Nevertheless, when she left the room she still learned something—she learned what red *looks like*. That is, she learned about the quale of redness. It follows that the quale of redness can't be a physical property. For if the quale of redness were a physical property, Mary would have known all about it when she was in the room: but she did not know all about the quale of redness when she was in the room, so the quale of redness isn't a physical property. Now physicalism claims that all properties are physical properties. So if the quale of redness isn't a physical property then physicalism is false. This is Frank Jackson's 'knowledge argument' against physicalism (Jackson 1982; 1986). So famous is the knowledge argument that it has even figured in a novel by David Lodge (2001).

I have just presented the knowledge argument in the way it is usually presented. Sometimes, though, Jackson varies the presentation slightly. The variation is important because it allows Jackson to defeat a significant objection which has been raised against the knowledge argument. The variant form of the argument begins the same way, with Mary learning all the physical facts of relevance to human color vision whilst trapped in a black and white room. However, in the variant, when Mary is released she exclaims, 'Gosh! Now I know what *everybody else's* visual experiences are like.' When she was in the room Mary had no idea what the visual experiences of the people outside the room were like. She did not know, for example, what it was like for them to see a ripe tomato in good light. But now that she has left the black and white room and seen a ripe tomato herself, she knows considerably more about the visual experiences of other people. Once again it follows that physicalism is false: there must be nonphysical properties of which Mary had no knowledge when she was in the black and white room. (I am assuming here that the people outside the room all have normal color vision: Mary *did know* what the visual experiences of colorblind people were like before she left the black and white room.)

Given the importance of the variant presentation to defending the knowledge argument, it's worth quoting Jackson's own text:

1. Mary (before her release) knows everything physical there is to know about other people.
2. Mary (before her release) does not know everything there is to know about other people (because she *learns* something about them on her release).

Therefore,

3. There are truths about other people (and herself) which escape the physicalist story.

(Jackson 1986: 293)

Before moving on it is worth briefly dealing with a common worry about the knowledge argument. The worry focuses on the sheer implausibility of the story

about Mary. If we really tried to conduct the Mary experiment, we would run into all sorts of practical problems. For example, if Mary cut herself she would immediately see red; if she had brown hair she would (eventually) see brown; if she bruised herself she would see blue. We could, of course, modify the story about Mary in various ways to avoid these problems. (Try this: Mary is forced to wear a special pair of glasses which make everything look black and white. Eventually the glass are removed and she exclaims . . .) However, the crucial point is that Jackson has offered us a *thought experiment*, and one of the great virtues of thought experiments is that they sidestep troublesome details like these.

Thought experiments can be very powerful. Amongst the most skilful users of thought experiments were the physicists Galileo and Einstein. Compared with Einstein's famous tram ride thought experiment, Jackson's Mary case is—well—pedestrian. In Einstein's thought experiment one of the laws of physics is set aside—the tram is pictured as traveling at the speed of light which, by Einstein's own theory, is physically impossible. So we can only condemn Jackson for using a thought experiment if we are also going to charge Galileo and Einstein with the same crime. And before we do *that* we'd better have some pretty convincing arguments against the use of thought experiments up our sleeve. (For a terrific discussion of thought experiments see Sorensen 1992.)

On the basis of the knowledge argument, Jackson decided that physicalism was false. What kind of position did he propose to put in its place? How did he conceive of the relationship between qualia and the brain?

In Section 1.4 we discussed property dualism. According to property dualism, mental states are nonphysical properties of the brain. One special kind of property dualism is *epiphenomenalism* (again see Section 1.4). According to epiphenomenalism, the physical properties of the brain cause a variety of nonphysical mental properties, but not vice versa. Jackson endorsed a special kind of epiphenomenalism. Whilst he rejected the view that *all* mental properties are epiphenomenal, he accepted that *qualia* are epiphenomenal. That is, he accepted the claim that although the physical properties of our brains give rise to nonphysical qualia, the qualia in turn do nothing—they are causally inert.

This is, to put it mildly, a strikingly counterintuitive view. It's natural to think that the hurt I experience when I stub my toe *causes* me to say 'ouch' and rub the damaged part. However, if qualia are indeed epiphenomenal, the hurt I experience when I stub my toe has no causal consequences whatsoever. In particular, I don't rub my toe because it hurts!

How did Jackson arrive at this extraordinary position? Since Jackson accepted the conclusion of the knowledge argument, he became a property dualist about qualia; that is, he accepted that qualia are nonphysical properties of the brain. Notice that if color qualia are indeed nonphysical properties of the brain, then

it is no surprise that Mary had no knowledge of them when she was in the black and white room. However, Jackson also accepted what I called in Chapter 1 the *explanatory completeness of physiology* (see Sections 1.3 and 1.4). That is, he accepted that human behavior can be fully explained by appealing to purely physical states such as muscle contractions and nerve impulses. Consequently, whilst he endorsed the claim that qualia are nonphysical, he could not allow them to impact upon behavior because if they did the explanatory completeness of physiology would be violated. So he adopted epiphenomenalism about qualia.

When discussing Jackson's epiphenomenalism about qualia I was careful to use the past tense. I did so because Jackson no longer accepts the conclusion of the knowledge argument, and no longer endorses epiphenomenalism about qualia. He has advanced several considerations against the knowledge argument, one of which we will look at in the next section.

1. Before she leave, She know all the fact
2. learn new fact when Mary leave the room. ie. what it is like to see red.
3. some facts aren't physical
4. Therefore, physicalism is false.

12.2 Responding to the knowledge argument

The knowledge argument has spawned a large literature, and a number of objections to the knowledge argument have been raised. In what follows I will consider four objections. My choice is not (I hope!) idiosyncratic. I have included the 'knowledge-how' reply due to Laurence Nemirow (1980) and David Lewis (1983*b*; 1990) because it's probably the best-known reply; a reply by Paul Churchland (1985; 1989) because it raises an important question; an especially interesting argument by Jackson and his co-author David Braddon-Mitchell (Braddon-Mitchell and Jackson 1996: 134–5); and a reply (better: a series of related replies) which insists that Mary could not have learned all the physical facts in the black and white room (see, for example, Horgan 1984; Loar 1990).

1. ***The knowledge-how reply.*** Philosophers distinguish two importantly different sorts of knowledge: knowledge-*that* and knowledge-*how*. Here are two examples of knowledge-that:

(i) Bloggs knows that Mt Everest is 8,848 meters high.

(ii) Bloggs knows that Central Park is in New York.

In both cases Bloggs knows a *fact* about the world. In the first case he knows the height of a certain mountain; in the second he knows the location of a certain park. (Notice that the locution 'knows that' appears in both sentences.)

Now consider the following two examples:

(iii) Bloggs knows how to play the trombone.

(iv) Bloggs knows how to swim.

objections
- P1 is false, Mary does not know all of the physical facts
- why? Language is limited.

- Two kind of knowledge
- Propositional → knowledge of facts or knowledge that
- Ability → skill knowledge or knowledge how

Premise 2. x. not learn new facts but new skill

Sentences (iii) and (iv) do not attribute to Bloggs knowledge of some fact or other; rather, they attribute to him a *skill*. Sentence (iii) attributes to Bloggs the skill of trombone playing; sentence (iv) attributes to him the skill of swimming. (Notice that the locution 'knows how' appears in both sentences.)

Let's now apply the knowledge-that/knowledge-how distinction to the knowledge argument.

When I described Mary's predicament I said that when she was in the black and white room she knew all the physical facts of relevance to human color vision. Clearly, this is an example of knowledge-that. Mary is described as knowing all the *facts* about color vision. She knew *that* light is reflected from a range of objects; *that* the reflected light is focused by the lens of the eye onto the retina; *that* information passes from the retina to the brain via the optic nerve; and so on. But what kind of knowledge did Mary acquire when she left the black and white room and saw a red object for the first time?

If we say that Mary acquired knowledge-that then it seems physicalism is indeed false. For Mary is supposed to know all the relevant physical facts when she is inside the room, so if she came to learn more facts when she left the room then those facts must be nonphysical facts. But if there are nonphysical facts for Mary to learn, then physicalism is false.

What about the other option? Perhaps what Mary acquired when she left the black and white room was knowledge-how. On this view, when Mary learned what the quale of redness is like, she learned some new skills. Which skills might they be? Well, Mary learned how to recognize red objects just by glancing at them in good light, and she learned how to imagine red objects.

This is the knowledge-how reply to the knowledge argument. The proponents of the knowledge-how reply insist that Mary did not acquire any new *facts* when she left the room, thereby preserving physicalism. However, they accept that Mary acquired *something* when she left the room: she acquired some new skills. The knowledge-how reply aims to reject Jackson's antiphysicalist conclusion but at the same time to satisfy our strong intuition that Mary gained something when she left the room.

In reply, Jackson (1986) admits that Mary acquired new skills when she left the room, but he doubts that that is all she acquired. It is here that the variant presentation of the knowledge argument becomes important (see previous section). On the variant presentation, Mary learns something about other people when she see a red object for the first time: she learns what their color visual experiences are like. To some extent her new knowledge of other people is knowledge-how; for example, she can now imagine what their color experiences are like. However, Jackson insists that Mary *also* acquired knowledge-that.

Jackson offers the following brilliant argument in favor of his claim that Mary acquired knowledge-that when she left the black and white room

(Jackson 1986: 294). There's an ancient position in philosophy called 'skepticism about other minds'. The skeptic about other minds denies that there are good reasons for thinking that anybody else has a mind. I'm not going to consider the arguments offered for and against this view. Rather, let's suppose that when Mary sees a red object for the first time she initially thinks, 'Now I know what other people's visual experiences are like', but then she starts to have skeptical doubts as to whether those other people have minds at all. After carefully considering the arguments for and against skepticism about other minds, she decides that all those people really do have minds; in particular, she decides that all those people have phenomenally conscious experiences. But what have her careful considerations been about? Surely not her ability to imagine red or recognize red objects in good light. Her concern was whether she was justified in thinking that other people experience the same color qualia as she is now experiencing; that is, she is concerned whether she has got the *facts* about other people's visual experiences right. It follows—provided other people really do have minds!—that Mary acquired at least some knowledge-that when she left the black and white room.

2. ***An argument against dualism?*** Imagine for a moment that we are all convinced of the truth of property dualism, and that our best theory of the relationship between the physical properties of the brain and the nonphysical mental properties is theory X. Now consider Joseph who has lived since birth in a black and white room. He is incredibly smart and has learned, via black and white textbooks, both theory X and all the physical facts about human color vision. Finally, Joseph is released from the room and sees a ripe tomato in good light. 'Gee,' he says, 'So that's what red looks like.' It seems, then, that Joseph *learned* something when he left the black and white room. But if he learned something when he left the room, theory X plus all the physical facts about color vision can't be the whole story about color vision. Assuming that Joseph has the physical facts right, it follows that theory X must be inadequate in some way. Since 'theory X' is just the name we gave to our best understanding of property dualism, it follows that property dualism is inadequate in some way. This line of reasoning, due to Paul Churchland (1985; 1989), strongly suggests that there is something wrong with the knowledge argument. For it seems that the knowledge argument can be used to 'prove' both the inadequacy of physicalism *and* the inadequacy of property dualism. (Indeed, it seems that the knowledge argument can be used to demonstrate the inadequacy of just about any metaphysical theory of the mind.)

What should we conclude from Churchland's ingenious objection? I think that Churchland's objection invites us to think about exactly what we can expect *any* theory to tell us about qualia. The nature of scientific theories is a controversial issue in the philosophy of science. On one view, a theory consists of a set of

sentences (often expressed mathematically) about some aspect of the world. Whilst this conception of scientific theories may not be universally accepted, it's certainly the case that the lessons Mary received in the black and white room consisted of sets of sentences. But if we think of theories in these terms, then it seems highly implausible that possessing a theory of color vision will tell us what it is like to see a ripe tomato in good light. Why would we expect reading and understanding a lot of black and white sentences to give us the *experience* of red? On this view, the knowledge argument tells us something about the limitation of theories, whether they be physicalist theories or theories of some other sort; it doesn't tell us anything about the metaphysics of qualia.

Theories may be limited in the way I have described because of the way our minds are structured. David Lewis has a nice metaphor which we can use to illustrate the present point. (See Lewis 1983*b*. Note that Lewis uses it primarily in support of the knowledge-how reply discussed above.) Imagine a computer which stores a description of any geometric figure with which it is presented. For example, if it is presented with a circle 12 cm in diameter it will store the description:

A circle 12 cm in diameter.

Now imagine a device which stores information about any geometric figure with which it is presented by making and retaining a copy of the figure. If presented with a circle 12 cm in diameter it literally makes and stores a circle 12 cm across. As it happens, both the computer and the copy-and-store machine are built into the same box. However, whilst they share the same box, there are no interesting connections between them. In particular, the copy-and-store machine cannot make (say) a 10-cm square on the basis of a description it received from the computer. As Lewis remarks, 'We might be rather like that' (Lewis 1983*b*: 132). It might be the case that the part of the brain which learns and retains physical theories of color qualia is not connected in relevant ways to the part of the brain which generates color qualia. Indeed, Paul Churchland (1989) has provided evidence that the human brain *is* like that. If this is right, Mary's failure to anticipate what the experience of redness is like is no evidence that physicalism is false. It only shows that the physical theory is not available to the qualia generator.

3. ***The 'There must be a reply' reply.*** We saw at the end of Section 12.1 that Jackson endorsed epiphenomenalism about qualia. On the basis of the knowledge argument he accepted that qualia are nonphysical properties of the brain; that is, he accepted a restricted version of property dualism. However, he had independent grounds for accepting the explanatory completeness of physiology. Consequently, he concluded that qualia have no causal impact on our behavior; that is, he thought that qualia are epiphenomenal.

We also saw at the end of Section 12.1 that Jackson no longer accepts the conclusion of the knowledge argument. With David Braddon-Mitchell he has advanced what they call the *'There must be a reply' reply*. They begin by noting that epiphenomenalism about qualia is a strikingly implausible doctrine. It entails, for example, that when I stub my toe I do not say 'ouch' and rub the injured part because it hurts, nor do I stop at the traffic lights because they look red. Braddon-Mitchell and Jackson argue that epiphenomenalism about qualia is *so* implausible that any doctrine which entails it must be wrong. (See Braddon-Mitchell and Jackson 1996: 134–5.)

Braddon-Mitchell and Jackson strengthen their case against epiphenomenalism about qualia by pointing out how very odd the Mary story becomes if we accept that qualia are epiphenomenal. According to the Mary story, when she leaves the black and white room and sees a red object for the first time she says, 'Wow! Now I know what red looks like'. It's natural to think that Mary said that because she had just experienced the quale of redness for the first time. That is, it's natural to think that Mary's exclamation was *caused* by her experience of redness. But if qualia are epiphenomenal they could not have caused Mary to say 'Wow!' because, by definition, epiphenomenal qualia don't cause anything.

It gets worse. The Mary story prompts the intuition that Mary learned something about qualia when she left the black and white room and saw her first red object. It's natural to think that she learned something about qualia because of the impact the qualia made on her; in particular, the qualia caused her to have certain beliefs. But if epiphenomenalism about qualia is true that can't be right. Consequently, the advocate of the knowledge argument is in the strange position of claiming that Mary acquires new knowledge of qualia when she leaves the room, but not in virtue of the causal powers of qualia. But if the qualia aren't responsible for Mary's new knowledge, how did it come about and why did she have to leave the black and white room to get it? Mysteries pile upon mysteries.

It's worth thinking about Braddon-Mitchell and Jackson's reply to the knowledge argument in a little more detail. Remember that Jackson came to believe that qualia are epiphenomenal because he believed *both* the conclusion of the knowledge argument *and* that physiology is explanatorily complete. It follows that if, as Braddon-Mitchell and Jackson urge, we reject epiphenomenalism about qualia, then we must reject either the conclusion of the knowledge argument or the explanatory completeness of physiology (or both). Braddon-Mitchell and Jackson are very reluctant to reject the explanatory completeness of physiology; consequently, they reject the conclusion of the knowledge argument.

For present purposes let's simply accept the explanatory completeness of physiology. Braddon-Mitchell and Jackson's argument comes down to this. The knowledge argument entails epiphenomenalism about qualia, but the latter

doctrine is crazy so something must be wrong with the knowledge argument. Braddon-Mitchell and Jackson candidly admit that they don't know *what's* wrong with the knowledge argument; their rejection of the knowledge argument is based solely on the fact that it entails epiphenomenalism about qualia. (It should be remembered that Jackson now believes he knows what's wrong with the knowledge argument.)

I find the 'There has to be a reply' reply convincing because, like Braddon-Mitchell and Jackson, I find epiphenomenalism about qualia close to unintelligible. I suspect, though, that antiphysicalists who are impressed by the knowledge argument are not going to be swayed by the 'There has to be a reply' reply. They will insist that they find the knowledge argument so convincing that they are prepared to accept what it entails—epiphenomenalism about qualia. They will insist that the knowledge argument has led us to a surprising discovery about the mind—that qualia are causally inert. There are no clear winners here.

4. *Did Mary know all the physical facts?* We are told that Mary knew all the physical facts when she was in the black and white room. But is that plausible? Might there be physical facts—in particular, facts about what it is like to see red—of which pre-release Mary would inevitably be ignorant? Considerations like this have prompted a variety of interrelated ways of replying to the knowledge argument.

We have already noted the suggestion that the power of theories might be limited so that not even grasping the true theory of the relationship between brain states and qualia would be sufficient to convey what it is like to see a ripe tomato in good light. In that case it may be that qualia are physical and yet Mary may not know, on the basis of her possessing all the relevant physical theories, what qualia are like. A second, related, proposal is that language is limited. On this view there are some physical facts—for example, facts about what it is like to see red—which simply cannot be conveyed in language. In order to know what it is like to see a red object in good light you have to have *seen a red object in good light*. No amount of talk substitutes for the experience itself. A third proposal is that there is something very special about the first person perspective. The facts we learn from the first person perspective are physical facts; nevertheless, we are accessing those facts in a way quite distinct from the third person perspective provided by science. Thus, whilst Mary knew all the relevant facts when she was in the black and white room, she only knew them from the third person, scientific, perspective. After she left the black and white room she became familiar with the same old facts in a new way (see Horgan 1984). And finally it has been proposed that there are certain concepts—*phenomenal concepts*—which can only be grasped by having the experience itself. The states which fall under these concepts are physical states picked out in a wholly new way. On this view, what Mary acquired when she left

the black and white room was a new set of concepts which allowed her to think about her (entirely physical) environment in a new way (Loar 1990).

My own view—for what it's worth—is that the correct reply to the knowledge argument lies somewhere in the terrain broadly indicated by the four ideas just sketched. However, filling in the details of these ideas is not easy. Why, for example, is language limited in this way? We seem to be able to express in language a great many facts, so why not facts about qualia? If a convincing reply to the knowledge argument is to be made along these lines we need a detailed account of language (or theories, or the first person perspective, or concepts) which makes it clear why the knowledge argument fails. It's fair to say that, at present, we have no such account.

Moreover, the antiphysicalist about qualia can press the following point. Here's why language is unable to express the facts about qualia—because qualia aren't physical! The asymmetry between the expressive power of language when it expresses facts about (say) ordinary physical objects, and its expressive power when it expresses facts about qualia, testifies to a substantial metaphysical difference between ordinary objects and qualia. A detailed theory of the limits of the expressive power of language may help us reply to the antiphysicalist, but as yet we lack such a detailed theory.

12.3 The explanatory gap

Modern scientists identify lightning with atmospheric electrical discharge. The identification of lightning with atmospheric electrical discharge explains the various features of lightning. For example, it explains how lightning can be harmlessly carried away from a building by a suitably placed copper wire (a lightning rod). Copper is much better at conducting electricity than is the concrete, masonry, and wood from which buildings are typically made. Consequently, the electrical discharge is quickly dissipated through the copper wire, leaving the building unharmed.

Now, according to physicalism, Bloggs's painful experience is identical to some state of his brain. However, unlike the identification of lightning with atmospheric electrical discharge, the identification of Bloggs's painful experience with a state of his brain seems to lack explanatory power. Say that Bloggs's painful experience is identified with the rapid firing of neurons in a certain part of his brain. How does that explain the painfulness? Why does that kind of firing in that part of Bloggs's brain hurt rather than, say, tickle? These questions do not seem to have answers. There is an *explanatory gap* between the firing in that part of the brain and the painfulness.

One of the principal proponents of the explanatory gap argument is the contemporary American philosopher Thomas Nagel. It was Nagel who introduced the phrase 'what it is like' as a way of drawing attention to phenomenally conscious experiences (see Section 11.1). In a famous paper he asked what it is like to be a bat (Nagel 1974). Bats find their way around in complete darkness by a process known as *echo-location*. They emit high-frequency 'squeaks' which echo off nearby objects, and then analyze the echoes in sophisticated ways to yield remarkably accurate representations of their surroundings. (Bat echo-location very closely parallels the sonar deployed by submarines. For a description of sonar, see Section 7.1.) So what is it like to be a bat? What is it like to build up an accurate representation of your surroundings using echoes? We humans have no idea—the experience is simply too far from our own. (Apparently some blind people have a very limited ability to echo-locate; however, there is no reason to assume that the phenomenally conscious experiences they have when they do so are anything like those of a bat.)

Notice that learning all the physical details about bats and how they echo-locate will not help you understand what it is like to be a bat. A lifetime spent studying the neurobiology of echo-location would no doubt reveal many interesting facts, but it would not move you any closer to understanding the phenomenal experiences of bats. There seems to be an unavoidable explanatory gap between studying the bat's brain and understanding its phenomenal life.

The contemporary American philosopher Joseph Levine has assembled two arguments in support of the claim that there is an explanatory gap between brain states and phenomenal experiences. He is not alone in advancing such arguments; however, his arguments are both significant in their own right and indicative of the general thrust of the literature in this area. In the remainder of this section I will briefly sketch Levine's arguments. (See Levine 1983; 1993; 2001. Related arguments have been advanced by Nagel 1974; Kripke 1980: Lecture III; and McGinn 1991; 1999.)

Levine begins by thinking about the case of water and H_2O. According to modern science, water is H_2O; that is, water is type identical to H_2O (for more on type identities see Section 3.1). The identification of water with H_2O is explanatorily satisfying in much the same way that the identification of lightning with atmospheric electrical discharge is explanatorily satisfying: it allows us to understand how water has the properties it in fact has. For example, once we realize that water is H_2O we can understand why water breaks down into hydrogen and oxygen gas in the process called 'electrolysis'.

Levine now shifts to the case of phenomenal consciousness. Let us take pain as our example, and let us assume that pain is identical to a physical property of the brain, P. We have seen that the various properties of water can be explained