Chapter 1

Consciousness and explanation

Martin Davies

1.1 Two questions about consciousness: 'what?' and 'why?'

Many aspects of our mental lives are conscious—an ache in tired muscles; the sight, smell, and taste of a glass of wine; feelings of happiness, love, anxiety or fear; trying to work out how best to test a hypothesis or structure an argument. It seems beyond dispute that at least some sensations, perceptions, emotional episodes, and bouts of thinking are conscious. But equally, there is much in our mental lives that is not conscious. It is a central idea in cognitive science that there can be unconscious information processing. It is also plausible that there can be unconscious thought and unconscious emotions; there are cases of 'perception without awareness'; and perhaps even bodily sensations can sometimes be unconscious.[1] What, then, is the difference between conscious and unconscious mental states? Is there, for example, something distinctive about the neural underpinnings of conscious mental states? An answer to this 'what?' question could be called (in some sense) an explanation of consciousness.

We might, however, expect rather more from an explanation of consciousness than just a principle or criterion that sorts conscious mental states from unconscious ones. Suppose that we were told about a neural condition, NC, that was met by conscious mental states but not by unconscious ones. Suppose that this was not just an accidental correlation. Suppose that the difference between meeting this neural condition and not meeting it really was the difference that

[1] Claims about unconscious thoughts and emotions are common in, but not restricted to, the psychoanalytic tradition. In ordinary life, it sometimes seems that we arrive at a solution to a problem by processes of thinking that do not themselves surface in consciousness, although their product does. For the conception of emotion systems as unconscious processing systems whose products are sometimes, but not always, available to consciousness, see LeDoux (1996, this volume). The term 'perception without awareness' is applied to a wide range of phenomena (Merikle *et al.* 2001) including blindsight (Weiskrantz 1986, 1997). For the proposal that unconscious mental states may include even sensations, such as pains, see Rosenthal (1991, 2005); for a recent discussion, see Burge (2007, pp. 414–419; see also 1997, p. 432).

makes the difference. There would remain the question *why* mental states that meet condition NC are conscious. Even if condition NC were to make the difference, it would not be *a priori* that it makes the difference. It would seem perfectly conceivable that condition NC might have been met in the absence of consciousness. So—we would ask—why, in reality, in the world as it actually is, is this condition sufficient for consciousness? The problem with this 'why?' question is that, once it is allowed as legitimate, it is apt to seem unanswerable.

The intractability of the 'why?' question is related to our conception of consciousness, a conception that is grounded in the fact that we ourselves are subjects of conscious mental states. The situation would be quite different if our conception of consciousness were a third-person conception, exhausted by structure and function—if it were a physical-functional conception. Our conception of a neurotransmitter, for example, *is* a physical-functional conception. There is a 'what?' question about neurotransmission: What chemicals make the difference? But, once we know the structure and function of GABA or dopamine, its role in relaying, amplifying, and modulating electrical signals between neurons, there is no further question why it is a neurotransmitter. That is just what being a neurotransmitter means. Similarly, if our conception of consciousness were a physical-functional conception then lessons about the nature of condition NC and about its role in the overall neural economy, about its constitution and connectivity, could persuade us that neural condition NC was consciousness—or, at least, that NC played the consciousness role in humans—because it had the right structure and function.

As things are, however, our conception of consciousness does not seem to be exhausted by structure and function and the 'why?' question remains. A neuroscientific answer to the 'what?' question would be of great interest but it would not render consciousness intelligible in neuroscientific terms. Consciousness would remain a brute fact. Between neural (or, more generally, physical) conditions and consciousness there is an *explanatory gap* (Levine 1983).

1.1.1 Positions in the philosophy of consciousness

Sometimes, the explanatory gap is presented as licensing a conclusion about the nature of reality itself, and not just about our conceptions of reality. It is argued that the existence of an explanatory gap between the physical sciences and consciousness supports the conclusion that consciousness is metaphysically or ontologically distinct from the world that the physical sciences describe. It would be no wonder that consciousness could not be explained in terms of the physical sciences if consciousness were something quite different from the physical world. The conclusion that consciousness falls outside the physical order is sometimes dramatized as the claim that there could, in principle, be a

creature physically just like one of us yet lacking consciousness—a zombie (Chalmers 1996; Kirk 2006), or even a complete physical duplicate of our world from which consciousness was totally absent—a zombie world. In line with this claim, David Chalmers proposes that 'a theory of consciousness requires the addition of *something* fundamental to our ontology, as everything in physical theory is compatible with the absence of consciousness' (1995, p. 210).

While Chalmers argues that consciousness is not wholly physical, it is more common (at least within academic philosophy) to assume or argue that some version of physicalism is true, so that consciousness must be part of the physical world (Papineau 2002).[2] Contemporary physicalists reject the duality of material and mental substances that Descartes proposed and also reject the duality of material and mental properties or attributes. According to physicalism, conscious mental states, processes, and events are identical to physical (specifically, neural) states, processes, and events. Furthermore, the phenomenal properties of conscious mental states (what being in those states is like for the subject) are the very same properties as physical properties of neural states or—if the claim of identity between phenomenal and physical properties seems too bold—the phenomenal properties are strongly determined by physical properties. The idea of strong determination in play here is that the phenomenal properties are necessitated by the physical properties. The phenomenal properties do not and could not vary independently of the physical properties; they *supervene* on the physical properties.[3]

Physicalist approaches to the philosophy of consciousness come in two varieties. Chalmers (1996) calls the two kinds of approach *type-A materialism* and *type-B materialism*. Some physicalists (type-A materialists) deny that there is an explanatory gap and maintain, instead, that consciousness can be fully and satisfyingly explained in physical terms. This option is, of course, mandatory for physicalists who agree with anti-physicalists like Chalmers that there is a good argument from the existence of an explanatory gap to the conclusion that consciousness falls outside the physical order.

Other physicalists (type-B materialists) allow that there is an explanatory gap but deny that there is a good argument from the gap to the anti-physicalist conclusion. In his development of the notion of an explanatory gap, Joseph Levine (1993) distinguishes two senses in which it might be said that the physical

[2] See Zeman (this volume, section 11.2.3) for some data about public understanding of the mind. In a survey of undergraduate students, '64% disputed the statement that "the mind is fundamentally physical"' (p. 294).

[3] We shall return (section 1.7.2) to the distinction between the strict version of physicalism (phenomenal properties are identical to physical properties) and the relaxed version (phenomenal properties are determined by, or supervene on, physical properties).

sciences *leave out* consciousness, the epistemological sense and the metaphysical sense. The claim that the physical sciences leave out consciousness in the epistemological sense is the claim that there is an explanatory gap. The claim that the physical sciences leave out consciousness in the metaphysical sense is the claim that consciousness falls outside the physical order. Levine says that the distinction between epistemological leaving out and metaphysical leaving out 'opens a space for the physicalist hypothesis' (1993, p. 126). Type-B materialist approaches typically involve two claims. First, the explanatory gap results from our distinctively subjective *conception* of consciousness. Second, there can be both scientific conceptions and subjective conceptions of the same physical reality (just as, in the familiar case of Hesperus and Phosphorus, there can be two concepts of a single object, the planet Venus). Type-B materialists maintain that there can be a duality of conceptions without a duality of properties.

1.1.2 Outline

This chapter begins with the subjective conception of consciousness that gives rise to the explanatory gap and the intractability of the 'why?' question. Next, there is a discussion of the approach to the study of consciousness that was adopted by Brian Farrell (1950), an approach that frankly rejects the subjective conception in favour of a broadly behaviourist one.[4] Farrell's approach serves as a model for subsequent type-A materialists.

The second half of the chapter is organized around Frank Jackson's *knowledge argument*—an argument for the anti-physicalist claim that phenomenal properties of conscious mental states are not physical properties. The knowledge argument 'is one of the most discussed arguments against physicalism' (Nida-Rümelin 2002) and the philosophical literature of the last 25 years contains many physicalist responses to the argument. Perhaps the most striking response is Jackson's own, for he now rejects the knowledge argument, adopting a type-A materialist approach and denying that there is an explanatory gap.

The type-A materialism that Jackson shares with Farrell denies that there is anything answering to our conception of consciousness to the extent that the conception goes beyond structure and function. In that respect, type-A materialism 'appears to deny the manifest' (Chalmers 2002, p. 251), and is probably the minority approach amongst philosophers who defend physicalism

[4] Brian Farrell was Wilde Reader in Mental Philosophy in the University of Oxford from 1947 to 1979. He died in August 2005, at the age of 93. The institutional and historical setting of the lecture on which this chapter is based (Oxford in the spring of 2006) invited extended reflection on Farrell's paper.

(although it is the approach adopted by such influential figures as Daniel Dennett and David Lewis). The more popular approach is type-B materialism, accepting that there is an explanatory gap but denying that this leads to the anti-physicalist conclusion (Chalmers 1999, p. 476): 'It simultaneously promises to take consciousness seriously (avoiding the deflationary excesses of type-A materialists) and to save materialism (avoiding the ontological excesses of the property dualist).' By considering the knowledge argument and responses to it, we shall be in a position to assess the costs and benefits of some of the most important positions in contemporary philosophy of consciousness.

1.2 The subjective conception of consciousness

We have distinguished two questions about the explanation of consciousness, the 'what?' question and the 'why?' question. The question *what* makes the difference between conscious and unconscious mental states seems to be a tractable question and many scientists and philosophers expect an answer in broadly neuroscientific terms—an answer that specifies the *neural correlates* of consciousness (Chalmers 2000; Block 2005; Lau, this volume). The question *why* this neuroscientific difference makes the difference between conscious and unconscious mental states is more problematic. As Thomas Nagel put the point over 30 years ago, in his paper, 'What is it like to be a bat?' (1974/1997, p. 524):

> If mental processes are indeed physical processes, then there is something that it is like, intrinsically, to undergo certain physical processes. What it is for such a thing to be the case remains a mystery.

Ned Block expressed a similar view in terms of qualia, the subjective, phenomenal, or 'what it is like' properties of conscious mental states (1978, p. 293):

> *No* physical mechanism seems very intuitively plausible as a seat of qualia, least of all a *brain*. ... Since we know that *we are brain-headed systems*, and that *we* have qualia, we know that brain-headed systems can have qualia. [But] we have no theory of qualia which explains how this is *possible*.

1.2.1 Nagel's distinction: subjective and objective conceptions

Nagel's announcement of mystery was not based on gratuitous pessimism about the progress of science but on an argument. The starting point was the thought that we cannot conceive what it is like to be a bat. The conclusion was that, although we can (of course) conceive what it is like to be a human, we cannot explain, understand, or account for (our) conscious mental states in terms of the physical operation of (our) brains. We should take a moment to

review the steps that led Nagel from the alien character of bat consciousness to the mystery of human consciousness.

The initial thought about bats can be extended to a distinction between two types of conception. We cannot conceive what it is like to be a bat and likewise a bat or a Martian, however intelligent, could not conceive what it is like to be a human—for example, what it is like to undergo the conscious mental states that you are undergoing now. These limitations reflect the fact that conceptions of conscious mental states as such are *subjective*; they are available from some, but not all, points of view. Roughly, the conscious mental states that we can *conceive* are limited to relatively modest imaginative extensions from the conscious mental states that we ourselves *undergo*. We cannot conceive what it is like to be a bat although we can conceive what it is like to be human. We cannot conceive what it is like for a bat to experience the world through echolocation although we can conceive what it is like for a human being to experience the red of a rose or a ripe tomato.

While grasping what a conscious mental state is like involves deployment of subjective conceptions, the physical sciences aim at objectivity in the sense that the conceptions deployed in grasping theories in physics, chemistry, biology, or neuroscience are accessible from many different points of view. The physical theories that we can grasp are limited, not by our sensory experience, but by our intellectual powers; and the conceptions that are required are, in principle, no less available to sufficiently intelligent bats and Martians than to humans.

1.2.2 Knowing what it is like

Nagel said (1974/1997, p. 521; emphasis added), 'I want to *know what it is like* for a bat to be a bat', and he went on to point out that the expression 'knowing what it is like' has two different, though related, uses (*ibid*. p. 526, n. 8; see also Nida-Rümelin 2002, section 3.3). In one use, knowing what a particular type of experience is like is *having a subjective conception* of that type of experience. There is a partial analogy between knowing what a type of experience is like and the 'knowing which' that is required for thought about particular objects (Evans 1982).[5] Knowing what a type of experience is like is similar to knowing

[5] Gareth Evans's (1982) theorizing about object-directed thoughts was guided by *Russell's Principle*, which says that in order to think about a particular object a thinker must *know which* object it is that is in question. Evans interpreted the principle as requiring discriminating knowledge, that is, the capacity to discriminate the object of thought from all other things. Initially, this may sound so demanding as to make object-directed thought an extraordinary achievement. But Evans's examples of ways of meeting the 'knowing which' requirement make it seem more tractable.

which object is in question in being a kind of discriminatory knowledge. In the case of thought about particular objects, there are many ways of meeting the 'knowing which' requirement: for example, presently perceiving the object, being able to recognize it, or knowing discriminating facts about it. In the case of thought about types of experience, there may also be many ways of meeting the 'knowing which' requirement. Having a subjective conception of a type of experience is meeting the 'knowing which' requirement in virtue (roughly) of being the subject of an experience of the type in question. (We shall refine this shortly.)

In a second use, knowing what it is like is *having propositional knowledge* about a type of experience, conceived subjectively. It is not easy to provide a philosophical account of having a conception or concept, but a subject who has a conception of something has a cognitive capacity to think about that thing. A subject who has a conception of a type of experience can deploy that conception in propositional thinking and may achieve propositional knowledge about that type of experience. He might know that he himself is having an experience of that type, or that he has previously had such an experience; and he may know something of the circumstances in which other people have experiences of that type. In the latter case, the subject knows what it is like for people to be in those circumstances.

It is plausible that a subjective conception of a type of experience can be deployed in thought even when the subject is not having an experience of the type in question. If that is right, then it must be possible for a subject to meet the 'knowing which' requirement in respect of a type of experience without concurrently being the subject of an experience of that type. On some accounts of having a subjective conception, it might be that remembering being the subject of an experience of the type in question would be sufficient to meet the 'knowing which' requirement. (Perhaps having a veridical apparent memory would suffice.) Alternatively, it might be proposed that meeting the 'knowing which' requirement involves being able to imagine being the subject of an experience of the type in question or being able to recognize other token experiences of which one is the subject as being of the same type again. (We shall return to these abilities to remember, imagine, and recognize in section 1.10.2.)

Michael Tye (2000) suggests that there are two different ways in which a subject can meet the requirements for having a subjective conception of a type of experience. In the case of a relatively coarse-grained experience type, such as the experience of red, a subject might meet the 'knowing which' requirement on the basis of long-standing abilities to remember, imagine, and recognize experiences of that type. In the case of a very fine-grained experience type, such as the experience of a specific shade of red, the limitations of

human memory may prevent a subject from reliably discriminating later experiences of that precise type from others. Nevertheless, it seems that a subject who is actually having an experience of that shade of red (and whose attention is not occupied elsewhere) has a subjective conception of that fine-grained experience type and knows what it is like to experience that specific shade of red. In such a case, the subject meets the 'knowing which' requirement in virtue of being the subject of an experience of the fine-grained type in question even if possession of the subjective conception lasts no longer than the experience itself.

1.2.3 Nagel's conclusion: physical theories and the explanation of consciousness

With these two uses of 'knowing what it is like' in mind, we can distinguish two claims that are immensely plausible in the light of Nagel's distinction between subjective and objective conceptions. The first claim is that subjective conceptions cannot be constructed from (are not woven out of) the objective conceptions that are deployed in grasping theories in the physical sciences. A subject might be able to deploy all the objective conceptions needed to grasp physical theories about colour vision without having any subjective conception of the experience of red. The second claim that is plausible in the light of Nagel's distinction is that there is no *a priori* entailment from physical truths to truths about conscious mental states conceived subjectively.

The second claim is not an immediate consequence of the first (Stoljar 2005; Byrne 2006) because *a priori* entailment of subjective truths by physical truths does not require that subjective conceptions should be constructible from physical conceptions. The second claim says that a subject who was able to deploy objective conceptions of physical states and who also *possessed the subjective conception of a particular type of experience* would not, just in virtue of having those conceptions, be in a position to know that a person in such-and-such a physical state in such-and-such a physical world would have an experience of that particular type.

If these claims are correct then physical theories, to the extent that they achieve the objectivity to which science aspires, will not say anything about conscious mental states conceived subjectively. We know what it is like to undergo various conscious mental states, but the conceptions that constitute or figure in that knowledge have no place in our grasp of objective physical theory. Nor will the content of our distinctively subjective propositional knowledge about conscious experience be entailed *a priori* by physical theory.

Once we grant the contrast between subjective conceptions and the objective conceptions that are deployed in grasping physical theories, the conclusion

of Nagel's argument is compelling. We cannot explain conscious mental states *as such*—that is, conceived subjectively—in terms of the physical operation of brains conceived objectively.

In a similar spirit to the Nagelian argument, Colin McGinn says (2004, p. 12):

> any solution to the mind-body problem has to exhibit consciousness as *conservatively emergent* on brain processes: that is, we must be able to explain how consciousness emerges from the brain in such a way that the emergence is not *radical* or *brute*.

And (*ibid.*, p. 15):

> What the theory has to do is specify some property of the brain from which it follows *a priori* that there is an associated consciousness …. *A priori* entailments are what would do the trick.

But *a priori* or conceptual entailments will not be available precisely because of the 'vastly different concepts' (p. 19) that figure, on the one hand, in the physical sciences of the brain and, on the other hand, in our knowledge of what it is like to undergo conscious mental states.

1.2.4 Subjective conceptions and physicalism

According to Nagel's argument, the explanatory gap is a consequence of the distinction between subjective conceptions and the objective conceptions that are deployed in grasping physical theories. On the face of it, this duality of conceptions is consistent with the truth of physicalism and, indeed, at the end of his paper, Nagel says (1974/1997, p. 524): 'It would be a mistake to conclude that physicalism must be false.'

If physicalism is true and conscious mental states fall within the physical order then they are part of the subject matter of objective physical theory. Similarly, if thinking about things, or conceiving of things, falls within the physical order then the activity of deploying conceptions—even deploying subjective conceptions—is part of the subject matter of objective physical theory. Thus, when we grasp physical theories by deploying objective conceptions, we may think about a physical event or process that is, in fact, the deployment of a subjective conception. But this does not require us to be in a position, nor does it put us into a position, to deploy that subjective conception ourselves. Even on a physicalist view of what there is in the world, grasping physical theories is one thing and deploying subjective conceptions is another. (In sections 1.11 and 1.12, we shall consider arguments that this duality of conceptions is not, in fact, consistent with physicalism.)

Tye argues that the explanatory gap presents no threat to physicalism because, really, there is no gap (1999/2000, p. 23): 'it is a cognitive illusion'. By claiming that there is no gap, Tye does not mean that there really are *a priori*

entailments from physical truths to truths about conscious mental states conceived subjectively. He agrees with Nagel that the distinction between subjective and objective conceptions guarantees that there are no such entailments. But he argues that it is a mistake to describe the absence of such entailments as a *gap* (*ibid.*, p. 34):

> [T]he character of phenomenal [subjective] concepts and the way they differ from third-person [objective] concepts conceptually guarantees that the question [*why* it is that to be in physical state P is thereby to have a feeling with this phenomenal character] has no answer. But if it is a conceptual truth that the question can't be answered, then there can't be an explanation of the relevant sort, *whatever* the future brings. Since an explanatory *gap* exists only if there is something unexplained that needs explaining, and something needs explaining only if it can be explained (whether or not it lies within the power of *human beings* to explain it), there is again no gap.

There are at least two important points to take from this bracing passage. First, if there are distinctively subjective conceptions of types of experience then there will be truths about conscious experience that are not entailed *a priori* by physical truths. So, a philosopher who maintains that all truths about conscious experience *are* entailed *a priori* by physical truths (a type-A materialist) must deny that there are distinctively subjective conceptions of the kind that Nagel envisages. Second, the absence of *a priori* entailment from physical truths to truths about conscious experience (subjectively conceived) is conceptually guaranteed (Sturgeon 1994). So, it is not an absence that will be overcome by progress in the physical sciences.

I shall not, myself, put these important points in Tye's way. Instead of saying, with Tye, that there is no explanatory gap, I shall say that there is an explanatory gap if there is no *a priori* entailment from physical truths to truths about conscious mental states conceived subjectively. The difference from Tye is terminological. I am prepared to allow that an explanatory gap exists even though what is unexplained is something which, as a matter of conceptual truth, cannot be explained.

1.3 Farrell on behaviour and experience: Martians and robots

In discussions of Nagel's (1974) paper, it is often noted that the 'what it is like' terminology and, indeed, the example of the bat, occurred in a paper by Brian Farrell, 'Experience', published in the journal *Mind* in 1950. I shall come in a moment to the use that Farrell made of the bat example. Before that, I need to describe the problem that Farrell was addressing—a problem which, he said, troubled physiologists and psychologists, even if not 'puzzle-wise professional philosophers' (1950, p. 174).

The problem is that scientific accounts of 'what happens when we think, recognize things, remember, and see things' *leave something out*, namely, the experiences, sensations, and feelings that the subject is having (*ibid.*, p. 171).[6] The experimental psychologist, for example, gathers data about a subject's 'responses and discriminations', dealing with 'behaviour' but not with 'experience'. Thus (p. 173): 'while psychology purports to be the scientific study of experience,… the science, in effect, does not include experience within its purview'. The problem that troubled the physiologists and psychologists was, in short, that the sciences of the mind leave out consciousness.

Farrell argues that there is really *no such problem* as the physiologists and psychologists take themselves to face. He asks us to consider the sentence (1950, p. 175):

> If we merely consider all the differential responses and readinesses, and such like, that X exhibits towards the stimulus of a red shape, we are leaving out the experience he has when he looks at it.

He argues that this is quite unlike ordinary remarks, such as:

> If you merely consider what Y says and does, you leave out what he really feels behind that inscrutable face of his.

The difference between the two cases is said to be this (p. 176): 'What we leave out [in the second sentence] is something that Y can tell us about [whereas] what is left out [in the first sentence] is something that X cannot in principle tell us about'. But why is it that X cannot tell us about what seems to be left out by a description of responses and readinesses, namely, his experience? Farrell answers (*ibid.*):

> He has already given us a lengthy verbal report, but we say that this is not enough. We want to include something over and above this, *viz.*, X's experience. It is useless to ask X to give us further reports and to make further discriminations if possible, because these reports and discriminations are mere behaviour and leave out what we want.

A critic of Farrell's argument might object at this point. For, even granting that X's report itself would be a piece of behaviour, it does not yet follow that what X would tell us *about* would be mere behaviour. On the contrary, it seems that X might tell us about the phenomenal properties of the experience that he had when presented with a red shape. So we need to be provided with a reason why X's apparent description of an experience should not be taken at face value.

[6] Farrell does not distinguish between epistemological and metaphysical 'leaving out' claims.

A major theme in Farrell's argument is the apparent contrast between 'behaviour' and 'experience', in terms of which the problem is raised. Farrell points out that, in ordinary unproblematic cases where behaviour is contrasted with experience, the term 'behaviour' is restricted to *overt* behaviour. But in the case of the putatively problematic contrast—where the sciences of the mind are supposed to leave out experience—the notion of behaviour is stretched to include 'the covert verbal and other responses of the person, his response readinesses, all his relevant bodily states, and all the possible discriminations he can make' (p. 177). Farrell insists that, once the notion of behaviour is extended in this way, we cannot simply assume that it continues to *contrast* with experience rather than *subsuming* experience. This theme is developed in discussion of two classic philosophical examples, Martians and robots.

1.3.1 Wondering what it is like: Martians, opium smokers, and bats

In the example of 'the man from Mars' (1950, p. 183), Farrell asks us to imagine that physiologists and psychologists have found out all they could find out about a Martian's sensory capacities and yet they still wonder what it would be like to be a Martian. He says that the remark, 'I wonder what it would be like to be a Martian', seems to be sensible because it superficially resembles other remarks, such as 'I wonder what it would be like to be an opium smoker' and 'I wonder what it would be like to be, and hear like, a bat' (*ibid.*).

If, in an ordinary *unproblematic* context, I wonder what it would be like to be an opium smoker, then I may suppose or imagine that I take up smoking opium and that I thereby come to learn how the addiction develops, for example. What I would learn in the hypothetical circumstances of being an opium smoker might, Farrell says, outrun what could be learned by the 'clumsy' scientific methods available at a given time. But it would not be different in principle from what could be learned from third-person observation. Thus (pp. 172–3):

> Quite often [a psychologist] places himself in the role of subject. … What is important to note is that by playing the role of observer-subject, he does not add anything to the discoveries of psychological science that he could not in principle obtain from the observation of X [another subject] alone.

According to Farrell, what I would learn about the experience of the opium smoker from the point of view of the observer-subject would not fall under the term 'behaviour' in the sense restricted to overt behaviour, but it would fall under the term in its extended sense that includes covert responses, response readinesses, discriminations, and so on.

In a similar way, I could unproblematically wonder what it would be like to be a bat. I could suppose that a witch turns me into a bat and that, from the privileged position of observer-subject, I learn something about the being's discriminations and response readinesses. But, on Farrell's view, if I were to spend a day or so as a bat then what I would learn would not outrun developed bat physiology and psychology. And he is quite explicit that it would require no distinctively subjective concepts or conceptions (p. 173):

> [N]o new concepts are required to deal with what [the psychologist's] own subject-observation reveals which are not also required by what was, or can be, revealed by his [third-person] observation of [another subject].

In this unproblematic kind of wondering what it is like to be an opium smoker or a bat, what I would learn about would be covert responses and internal discriminations, behaviour in the extended and inclusive sense of that term. This would also be the case if I unproblematically wondered what it would be like to be a Martian (p. 185): 'the "experience" of the Martian would … be assimilable under "behaviour" '.

The example of the Martian began, however, with a kind of wondering that was supposed to be quite different from this unproblematic wondering about behaviour in the inclusive sense of the term. It was supposed to be a *problematic* wondering about something that would inevitably be left out by the sciences of the Martian mind—a wondering about experience as contrasted, not only with overt behaviour, but even with behaviour in the extended and inclusive sense of the term. Farrell's point is that, while unproblematic wondering is 'sensible', this putatively problematic wondering is 'pointless' (p. 185). We have no right to assume that this contrast—between experience and behaviour in the inclusive sense—is legitimate.

A critic of Farrell's argument might concede this point but also insist on another. We cannot simply assume that behaviour in the inclusive sense *contrasts* with experience; but equally we cannot simply assume that it *subsumes* experience. Until we have a positive argument for subsumption, the relationship between behaviour and experience should remain an open question. We shall come to Farrell's positive arguments shortly (section 1.4); but, before that, we review the second of the two classic philosophical examples, the robots.

1.3.2 Robots—and the criteria for having a sensation

The question under discussion in the example of the robot is whether we need to retain the contrast between behaviour and experience in order to say (1950, p. 189): 'If a robot were to behave just like a person, it would still not have any sensations, or feelings.'

Farrell's answer to the question comes in two stages. First, in ordinary talk about robots, the unproblematic contrast between experience and overt behaviour is adequate for the purpose. A robot, in the ordinary sense of the term, duplicates the overt behaviour of a human being but not the covert responses, bodily states, internal discriminations, and so on. So, second, if the example of the robot is to present a problem for Farrell's view then we must be thinking of a robot that duplicates, not only our overt behaviour, but all our covert responses and internal discriminations as well. But then, Farrell says, he has already argued that we cannot presume upon a contrast between experience and behaviour in this extended and inclusive sense.

In order to avoid the 'muddle' that results, according to Farrell, from this 'unobserved departure from the ordinary usage of "robot"' (p. 190), we could set aside that term for the time being. Then there are two ways that we might describe a mechanical system that duplicates the overt and covert, external and internal, behaviour of a person. On the one hand, we might allow, in line with what Farrell regards as our 'usual criterion' for having a sensation, that the mechanical system has sensations. On the other hand, we might adopt a more demanding criterion for having a sensation and deny that the mechanical system has sensations on the grounds that it is not a living thing.

Does either way of describing the mechanical system present a problem for Farrell's view about experience? If, on the one hand, we allow that a system that produces the right external and internal behaviour has experience then clearly the example provides no reason to retain a contrast between experience and behaviour in the inclusive sense. If, on the other hand, we insist that, while mechanical systems produce behaviour, only a living thing has experience then, of course, we do retain a kind of contrast between experience and behaviour in the inclusive sense. This more demanding criterion allows us to deny experience to inanimate robots. But the contrast between mechanical systems and living things has no relevance to questions about the mental lives of human beings, Martians, or bats. Farrell thus concludes that the example of the robot does not present a problem for the *behaviourist psychology of organisms*.

Bringing his discussion of robots even closer to contemporary philosophy of consciousness, Farrell invites us to consider a series of imaginary examples of robots that duplicate our external and internal behaviour and are increasingly like living things. He suggests that, as we progress along this series, it will be increasingly natural to allow that the robots have experience—sensations and feelings: (p. 191):

> General agreement [to allow the attribution of experience] would perhaps be obtained when we reach a machine that exhibited the robot-like analogue of reproduction, development and death.

Farrell's position thus leaves no conceptual space for zombies. It licenses the attribution of experience to a hypothetical living thing that duplicates our overt behaviour, covert responses, internal discriminations and bodily states. Consciousness is entailed *a priori* by life plus the right behaviour.

1.4 Experience from the third-person point of view

I have described Farrell's view that experience is subsumed by behaviour and have indicated some of the ways in which Farrell defended his position against the objection that we need the distinction between behaviour and experience in order to say the things that we want to say about Martians and robots. But it would be reasonable to ask what considerations motivated Farrell's view in the first place.

Part of the answer is that Farrell regarded scientists' concerns about consciousness as manifestations of their 'occupational disease of traditional dualism' (p. 170)—the dualism against which Gilbert Ryle argued in *The Concept of Mind* (1949). The conscious mind, as conceived by the dualist, was supposed to fill what would otherwise be gaps in causal chains. It was supposed to provide the middle part of a causal story that begins with physical processes leading from stimulation of sensory surfaces and ends with physical processes leading to contractions of muscles. As against this dualism, Farrell argued that the causal story leading all the way from sensory stimulation to overt behaviour could be told in terms of factors that, aside from being covert and internal rather than overt and external, could be grouped with behaviour—causal factors such as covert responses, discriminations, response readinesses, and bodily states.

There are also more specific points that figure in the motivation for Farrell's view. I consider two: Farrell's claim that experience is featureless and his rejection of distinctive first-person knowledge of experience.

1.4.1 Featureless experience

Immediately after introducing the apparent contrast between behaviour and experience, Farrell argues that experience, if it is contrasted with behaviour in the extended and inclusive sense, is 'featureless' (1950, p. 178). We are to consider X in the role of observer-subject looking at a red patch and ask whether there is anything about X's experience that he can discriminate. Farrell's answer is that there is not (*ibid.*):

> If he does discriminate something that appears to be a feature of the experience, this something at once becomes, roughly, either a feature of the stimulus in the sort of way that the saturation of the red in the red shape is a feature of the red shape, or a feature of his own responses to the shape. X merely provides us with further information about the behaviour that he does and can perform.

Here, Farrell presents two options for what we might be tempted to regard as a discriminated feature of an experience. Either it becomes a feature of the worldly stimulus or else it becomes a feature of the subject's response (that is, the subject's *behavioural* response, in the inclusive sense of the term).

The first option is that a putative feature of experience is better conceived as a feature of the worldly stimulus. David Armstrong (1996) takes this as an anticipation of the *representationalist* proposal that the phenomenal properties of an experience are determined by its representational properties—that is, by how it represents the world as being. I shall consider representationalism later (section 1.9). For now, let us note that a critic of Farrell's argument might ask how the view that experiences have representational properties is supposed to be consistent with the claim that experiences are featureless. For, intuitively, how an experience represents the world as being is an aspect of what it is like for a subject to undergo that experience, an aspect or feature of its phenomenology.

A critic might also have a worry about the second option in the quoted passage, the idea that the discriminated feature of an experience becomes a feature of the behavioural response. The critic might urge that it is not obvious how the fact that X's response *is* a piece of behaviour is supposed to support the claim that X's response provides information only *about* behaviour. (In essence, this is the same objection that was entered at an earlier point in Farrell's argument—see the beginning of section 1.3) We still need to be provided with a reason why X's behaviour should not be taken at face value, as evidence that he has discriminated a feature of his experience.

1.4.2 Acquaintance and the first-person point of view

Farrell himself anticipates an objection to his claim that experience is featureless, namely, that from the fact that experience has 'no features that can be described, or discriminated, or reported in a laboratory' it does not follow that experience has no features at all. He imagines an opponent saying (1950, p. 181):[7]

> [Experience] may still possess features with which we can only be acquainted. … When, for example, we look at a red patch, we all just *know* what it is like to have the corresponding experience, and we all just *know* how it differs from the experience we have when looking at a green patch.

[7] The imagined opponent's proposal is a striking anticipation of McGinn's comment (2004, p. 9): 'if we know the essence of consciousness by means of acquaintance, then we can just see that consciousness is not reducible to neural or functional processes (say)—just as acquaintance with the colour red could ground our knowledge that redness is not the same as greenness, say'.

He also has the opponent propose that the problem lies in restricting observation to the third-person case (p. 183). Farrell responds in his own person that experience remains featureless even if we allow first-person observation since apparent expressions of first-person knowledge about our experiences of worldly objects are really based on our discrimination of our responses to those objects (*ibid.*): 'we are … liable to mistake features of our responses to the [object] for some indescribable and ineffable property of the experience'.

At this stage, a critic might reckon Farrell's response to be unsatisfying, since there is still no direct argument against the idea of features of experience that can be discriminated from a first-person point of view. But, after the discussion of the example of the man from Mars, Farrell returns to first-person knowledge ('Knowing at first hand', p. 185). Here, the opponent is imagined to object that wondering what it is like to be a Martian is 'wondering what it would be like to have first-hand knowledge of the experience of a Martian' and that this first-hand knowledge would clearly be quite different from anything that one could learn by 'hearing a description'. We already know that Farrell is bound to reject this objection by insisting that the observer-subject learns about covert responses and internal discriminations and that this knowledge is available, in principle, to third-person observation and conception. But he now advances a new response.

Knowledge at first hand, in the ordinary use of the term, is contrasted with knowledge at second hand, which is learning from someone else. But in the case of knowing what it is like to be a Martian, Farrell's opponent envisages our knowing at first hand something that it is impossible to learn at second hand, knowing by acquaintance something that it is impossible to learn by description. So, in the problematic case as it is conceived by the opponent, knowing at first hand 'is not contrastable with anything [and so] this objection simply has not given a use to the expression "to know at first hand"' (p. 186).

Here Farrell makes use of a *contrast argument*, a kind of argument that was deployed by Ryle in *Dilemmas* (1954). Ryle says, for example (1954, p. 94): 'There can be false coins only where there are coins made of the proper materials by the proper authorities'; and (*ibid.*, p. 95): 'Ice could not be thin if ice could not be thick'. Similarly, Farrell is arguing that what could not be known at second hand could not be known at first hand.

Contrast arguments can sometimes be persuasive. For example, if thin ice is defined as ice that is thinner than average, then not all ice can be thin ice. If there is to be ice that is thinner than average then there must also be some ice that is thicker than average. But, in general, contrast arguments do not succeed in showing that if an expression does not apply to anything then a contrasting expression does not apply to anything either. A philosopher who claims that,

as a matter of necessity, there are no immaterial substances is not thereby saddled with the conclusion that there are no material substances either—nor with the conclusion that the expression 'material substance' has not been given a use.

In the case of Farrell's contrast argument, the expression:

(1) knows at first hand what it is like to be a Martian

contrasts with:

(2) knows at second hand what it is like to be a Martian.

The argument turns on the claim, made by Farrell's opponent, that, as a matter of necessity, expression (2) does not apply to anyone. Nobody can know at second hand what it is like to be a Martian. But—as is generally the case with contrast arguments—Farrell's argument does not succeed in showing that his opponent is saddled with the conclusion that expression (1) cannot apply to anyone either, nor with the conclusion that the opponent 'has not given a use' to expression (1).

1.5 **Farrell, Dennett, and the critical agenda**

More than 20 years before Nagel (1974), Farrell considered the question what it is, or would be, like to be a bat. But, as we have now seen, Farrell used the question for purposes that were completely opposed to the ideas in Nagel's paper. According to Farrell, the facts about experience do not outrun the facts that are available to the sciences of the mind by third-person observation and there are no distinctively subjective, first-person concepts that are deployed in our knowledge about experience. When physiologists and psychologists worry that their accounts are incomplete because they leave out experience, 'their fears are groundless' (1950, p. 197).

There are questions about experience that may seem to be problematic for Farrell's behaviourist account of consciousness—questions about what it would be like to be a bat or a Martian; about whether a robot could have experiences; about features of experience that a subject can discriminate; about acquaintance with phenomenal properties; and about distinctively first-person knowledge. But Farrell argues that these apparently problematic questions rest on various philosophical errors—on the unwarranted assumption that behaviour, in the inclusive sense, continues to contrast with experience rather than subsuming it; on the failure to apply usual criteria; on the assumption that experience itself has features that can be discriminated; on the confusion between features of our responses to worldly objects and phenomenal properties of experience; and on the failure to give meaning to the terms in which questions are cast.

Farrell's view is strikingly similar to the account of consciousness that Dennett (1988, 1991, 2005) has developed in recent years, although there is also a difference of dialectical context between them. The similarity is clear if we consider Farrell's insistence that there is no knowledge available to the observer-subject that is not also available to third-person observation (section 1.3.1) alongside Dennett's 'A third-person approach to consciousness' (2005, chapter 2), or Farrell's implied rejection of the conceivability of zombies (section 1.3.2) alongside Dennett's 'The zombic hunch: Extinction of an intuition?' (2005, chapter 1), or Farrell's rejection of indescribable and ineffable properties of experience (section 1.4.2) alongside Dennett's 'Quining qualia' (1988).

The difference of dialectical context is this. Farrell was addressing a problem that was raised by scientists—they feared that their accounts were bound to leave out experience. Farrell thought that philosophy could show that the scientists' fears were groundless. In contrast, Dennett regards himself as removing obstacles to progress towards a science of consciousness that have been erected, not by worried scientists, but by other philosophers—particularly, by philosophers who say that there is an explanatory gap.[8]

1.5.1 The need for a critical agenda

The first choice point in the philosophy of consciousness is whether to affirm or deny that there is an explanatory gap, that the physical sciences leave out consciousness in the epistemological sense, that there is no *a priori* entailment from physical truths to truths about conscious mental states conceived subjectively. Philosophers who deny that there is an explanatory gap (Dennett, Farrell) are able to proceed directly to type-A materialism. Those who allow that there is an explanatory gap (Block, Chalmers, Levine, McGinn, Nagel) face a second choice: type-B materialism or dualism.

We observed earlier that a type-A materialist must deny that there are distinctively subjective conceptions of the kind that Nagel envisages. As Chalmers (2002) notes, a type-A materialist may appear as a reductionist or as an eliminativist about consciousness, promoting a behaviourist or functionalist conception of consciousness or saying that there is no such thing as

[8] See the subtitle of his book, *Sweet Dreams* (2005), 'Philosophical obstacles to a science of consciousness', and the critical discussion of Block, Chalmers, Levine, McGinn, and Nagel, therein. My own view, in contrast, is that it does not obstruct progress towards a science of consciousness to point out that, if we have distinctively subjective conceptions of types of experience, then truths about conscious mental states conceived subjectively will not be entailed *a priori* by physical truths.

consciousness as it is conceived subjectively (nothing in reality corresponds to distinctively subjective conceptions). As Chalmers also points out, type-A materialism involves 'highly counterintuitive claims [that] need to be supported by extremely strong arguments' (2002, p. 251).

It is inevitable, then, that Farrell's argument develops a partly critical agenda. He rejects the very idea of distinctively first-person conceptions of types of experience; and he rejects the idea of a distinctive kind of knowledge gained by first-person acquaintance with the features of experience. On his view, such conceptions, and the apparently problematic questions about experience to which they give rise, are based on philosophical and conceptual errors. The proper conceptions of conscious mental states or types of experience can be constructed out of objective conceptions of behaviour in the inclusive sense. There is a corresponding critical agenda in Dennett's work. Just as Farrell argues that there are no discriminable features of experience with which subjects are acquainted, so Dennett argues that 'there are no such properties as qualia' conceived as 'directly and immediately apprehensible in consciousness' (1988, pp. 43, 47).

There remains, of course, a substantive question whether Farrell's critical agenda is effective, whether his negative arguments are sufficiently strong. At various points, I have noted ways in which a critic might respond to his arguments. More generally, most contemporary philosophers of consciousness would reject Farrell's apparent commitment to Rylean behaviourism and, particularly, his use of a contrast argument to cast doubt on the idea of knowing at first hand what a type of experience is like. A similar question can, of course, be raised concerning the critical aspect of Dennett's work.[9]

[9] I noted earlier (section 1.1.2) that type-A materialism—the approach adopted by Dennett—is probably the minority approach amongst philosophers of consciousness who defend physicalism. However, in this chapter I provide no details of Dennett's position. A proper assessment of the critical aspect of his work would need to consider the third-person approach to studying consciousness that he calls 'heterophenomenology' (the phenomenology of another; see Dennett 1991). The connection between the heterophenomenological approach and the more explicitly critical aspect of Dennett's work is apparent in the following passage in which Dennett criticises philosophers who assume that, in addition to recognitional and discriminatory capacities, there is 'a layer of "direct acquaintance" with "phenomenal properties"' (2007, p. 20): 'These [recognitional/discriminatory] capacities are themselves the basis for the (illusory) belief that one's experience has "intrinsic phenomenal character," and we first-persons have no privileged access at all into the workings of these capacities. That, by the way, is why we shouldn't do auto-phenomenology. It leads us into temptation: the temptation to take our own first-person convictions not as data but as undeniable truth.'

1.6 The knowledge argument

Type-A materialism, the kind of position adopted by Farrell and Dennett, is both conceptually and metaphysically reductionist. It is conceptually *deflationist* physicalism (Block 2002; Papineau 2002). The opposite position, dualism, is committed to both a duality of conceptions and a duality of properties. (A dualist may also be committed to a duality of states, processes, and events and perhaps—as in Descartes's case—a duality of substances.) Type-B materialism is an intermediate position, combining conceptual dualism with metaphysical reductionism. It is conceptually *inflationist* physicalism.

In recent philosophy of consciousness, Jackson's (1982, 1986) knowledge argument is one of two prominent attempts to argue from something like the explanatory gap or the epistemological 'leaving out' claim to the dualist conclusion that physicalism is false. The argument features Mary the brilliant scientist who, in her black-and-white room, learns everything about the physical world and then, on leaving the room for the first time, sees something red. The powerful intuition generated by the story of Mary is that, when she first sees a red rose or a ripe tomato she gains new knowledge. Now she knows, whereas before she did not know, what it is like to see red. Since Mary already knew all the physical facts, the knowledge argument concludes that there are facts that are not physical facts and that physicalism is therefore false.[10]

The other major argument for dualism in recent philosophy of consciousness is Chalmers's (1996) *conceivability argument*, which also begins from an epistemological premise. This argument proceeds from the premise that zombies are conceivable (zombies are not *a priori* impossible) to the metaphysical claim that zombies are possible (zombies exist in some possible world) and, thence, to the conclusion that physicalism is false (the phenomenal properties of conscious mental states are not strongly determined or necessitated by physical properties). The knowledge argument and the conceivability argument raise many of the same issues—particularly concerning the transition from epistemology to metaphysics—and they present philosophers of consciousness with the same options of type-A materialism, type-B materialism, and dualism. In the remainder of this chapter, the discussion of these options is organized around the knowledge argument.

[10] For a well-chosen sample of philosophical discussion of the knowledge argument, see Ludlow *et al*. (2004). The introduction by Stoljar and Nagasawa provides a helpful overview, as does the review of the book by Byrne (2006). The knowledge argument also plays a major role in David Lodge's novel *Thinks …* (2001).

1.6.1 The structure of the knowledge argument

It seems that, when Mary is released, she learns something new, something that she could not have worked out *a priori* from what she knew before she saw a red rose or a ripe tomato. This is an epistemological intuition that would support the claim that there is an explanatory gap—that is, the claim that the physical sciences leave out consciousness in the epistemological sense. But it is not immediately clear how it could justify the conclusion that physicalism is false—that is, the conclusion that the physical sciences leave out consciousness in the metaphysical sense.

Jackson makes the transition from epistemology to metaphysics by drawing on a crucial component of his overall philosophical position—a component that was not explicit in the earliest presentations (1982, 1986) of the knowledge argument. This is the claim that, if physicalism is true then there is an *a priori entailment* from the true physical story about the world to the true story about conscious mental states and their phenomenal properties.[11] Daniel Stoljar and Yujin Nagasawa explain this component of Jackson's position in terms of the *psychophysical conditional*, 'If P then Q', where P is the conjunction of all the physical truths and Q is the conjunction of all the psychological truths (2004, p. 15): '[W]e should assume that, in both the 1982 and 1986 essays, Jackson was supposing… that: if physicalism is true, the psychophysical conditional is *a priori*.'

In summary, we shall consider the knowledge argument as depending on two main premises. The first premise is epistemological:

(P1) Mary learns something new on her release.

The second premise is the principle linking epistemology and metaphysics:

(P2) If physicalism is true then the psychophysical conditional is *a priori*.

A powerful intuition supports the first premise (P1). On her release, Mary learns something that she could not have worked out *a priori* from what she knew in her black-and-white room, even though she already knew all the physical truths. If the first premise is true then the psychophysical conditional is not *a priori*. In that case, by the second premise (P2), physicalism is false.

[11] See Jackson (1995)—a postscript to Jackson (1986). For a more detailed account, see Jackson (1998a), Chapter 1, esp. pp. 6–14 and 24–7 on the entry by entailment thesis and Chapter 3, esp. pp. 80–3 on the question of *a priori* deducibility.

1.6.2 An objection to the second premise

The principle (P2) linking physicalism to *a priori* entailment certainly helps with the knowledge argument's transition from an epistemological first premise to a metaphysical conclusion. But the principle itself seems to be open to the objection that the metaphysical determination or necessitation of phenomenal facts by physical facts could be *a posteriori* rather than *a priori*. In Saul Kripke's (1980) famous example, it is a necessary *a posteriori* truth that water is H_2O. So the fact that H_2O covers most of the planet determines *a posteriori* that water covers most of the planet. Why cannot the *a posteriori* determination of facts about water by facts about H_2O serve as a model for the determination of phenomenal facts by physical facts?

Jackson responds to this objection by arguing that familiar examples of *a posteriori* determination, such as the example of water and H_2O, do not support the idea of *a posteriori* determination of phenomenal facts by the *totality* of physical facts.[12] The fact that water covers most of the planet is determined *a posteriori* by the fact that H_2O covers most of the planet; but it is determined *a priori* by a *richer* set of facts about H_2O. The reason is that, on Jackson's view, it is *a priori*—indeed, a matter of conceptual analysis—that water is whatever stuff is colourless, odourless, falls from the sky, and so on. It is *a priori* that water is whatever stuff 'fills the water role'.[13] Consequently, there is an *a priori* entailment from the facts that H_2O covers most of the planet *and that* H_2O fills the water role to the fact that water covers most of the planet.

In a similar way, Jackson says (1995/2004, p. 414):

> A partial story about the physical way the world is might logically necessitate the psychological way the world is without enabling an *a priori* deduction of the psychological way the world is. … But the materialist is committed to a complete or near enough complete story about the physical way the world is enabling in principle the *a priori* deduction of the psychological way the world is. … I think it is crucial for the truth of materialism (materialism proper, not some covert form of dual attribute theory of mind) that knowing a rich enough story about the physical nature of our world is tantamount to knowing the psychological story about our world.

[12] See Jackson (1995, 2003), Braddon-Mitchell and Jackson (2007, pp. 139–40).

[13] This part of Jackson's view would be roughly captured by saying that the term 'water' is a descriptive name with its reference fixed by the description 'the stuff that is colourless, odourless, falls from the sky, and so on' or by saying that the term 'water' behaves semantically and modally like the description 'the stuff that *actually* (that is, in the actual world) is colourless, odourless, falls from the sky, and so on'. See Davies and Humberstone (1980) for an early development of this view in the framework of two-dimensional semantics.

We shall return later (sections 1.8.2 and 1.8.3) to the issue of distinguishing physicalism from 'some covert form of dual attribute theory'. For the moment, I want to continue with the question how the second premise of the knowledge argument—the principle linking physicalism to *a priori* entailment—is to be motivated.

1.6.3 The second premise and type-A materialism

The second premise says, in effect, that if physicalism is true then type-A materialism is true. So long as both type-A and type-B materialism are options for the physicalist, the premise is open to the obvious objection that physicalism might be true without type-A materialism being true, because type-B materialism might be true. Now, according to type-B materialism, the determination of phenomenal facts by physical facts is *a posteriori* rather than *a priori*. So it is certainly relevant to point out, as Jackson does with the example of water and H_2O, that *a posteriori* determination by a partial set of facts may be consistent with *a priori* determination by a richer set of facts. But, although this is relevant, it does not yet go to the heart of the matter.

Both a type-A materialist and a type-B materialist will agree that, if our conception of water is a physical-functional conception, then water facts are entailed *a priori* by H_2O facts. This is not to say that all type-B materialists accept that we do have a physical-functional conception of water. For example, Brian McLaughlin says (2005, p. 280, n. 31): 'I regard that as an unresolved issue.' But it is certainly open to a type-B materialist to agree with Jackson that the state of water covering most of the planet *can* be explained in terms of facts about H_2O via functional analysis of the concept of water.

The type-B materialist disagrees with the type-A materialist, however, over the question whether water facts are relevantly similar to phenomenal facts and, particularly, whether a functional conception of water is a good model for our conceptions of types of experience. Nagel's account of the contrast between subjective conceptions and objective conceptions, and Levine's claim that there is an explanatory gap, both depend on our subjective conceptions of types of experience *not* being functional conceptions. Indeed, McLaughlin says (2005, p. 280): 'On one interpretation, [Levine's] *explanatory gap* thesis is the thesis that states of phenomenal consciousness cannot be *physically explained via … functional analysis*.'

A type-A materialist denies that there is an explanatory gap and denies that there are distinctively subjective conceptions of types of experience. Type-A materialism is conceptually deflationist physicalism. So a defence of the second premise of the knowledge argument against the objection that physicalism might be true without type-A materialism being true must go beyond the uncontested

example of water and H_2O. It must include an argument against the conceptually inflationist option of type-B materialism—an argument to show that subjective conceptions and the explanatory gap are inconsistent with physicalism.

1.6.4 The simplified knowledge argument

The second premise, (P2), of the knowledge argument—made explicit by Stoljar and Nagasawa (2004, p. 15)—appears to be motivated by the background assumption that conceptually deflationist physicalism (type-A materialism) is the only physicalism worthy of the name, so that type-B materialism can be rejected. The prospects for type-B materialism will be assessed later (sections 1.11 and 1.12). In the meantime, we can consider a simplified version of the knowledge argument from the epistemological first premise (as before):

(P1) Mary learns something new on her release.

and a second premise that is now true by definition:

(P2A) If type-A materialism is true then the psychophysical conditional is *a priori*.

to the conclusion that type-A materialism is false.

The simplified knowledge argument seems to be valid and the second premise (P2A) is true by the definition of type-A materialism. The argument presents us with three options. If we reject the conclusion and accept type-A materialism then we must also reject the epistemological premise (P1). If we accept the epistemological premise then we must also accept the conclusion, reject type-A materialism, and choose between type-B materialism (conceptually inflationist physicalism) and dualism.[14]

[14] The claim that the simplified knowledge argument presents just three options—type-A materialism, type-B materialism, and dualism—involves a degree of simplification. I assume that it is legitimate to include in the setting-up of the example that Mary already knows all the physical facts while she is in her black-and-white room. I also assume that, if Mary learns something new on her release, then she gains propositional knowledge. Each of these assumptions might be rejected, providing two more options (see Byrne 2006). According to the first option, there are physical facts of which the physical sciences tell us nothing (Stoljar 2001, 2006). According to the second option, Mary gains 'know how', rather than propositional knowledge, on her release. This is the ability hypothesis, discussed in section 1.10.2. Finally, the validity of the simplified knowledge argument might be challenged. Someone might deny that the first premise (P1) really entails that the psychophysical conditional is not *a priori*. As we observed earlier (section 1.2.3), the claim that there is no *a priori* entailment from physical truths to truths about conscious mental states conceived subjectively is not an immediate consequence of the claim that subjective conceptions cannot be constructed from objective conceptions. For discussion of this final option, see Byrne (2006), Nida-Rümelin (1995, 2002) and Stoljar (2005).

The simplified knowledge argument no longer depends on the assumption that type-B materialism can be rejected. The role of that assumption, if it could be justified, would be to license the transition from the limited conclusion of the simplified knowledge argument to the more sweeping conclusion of the original knowledge argument, namely, that physicalism is false.

1.7 The argument for physicalism

There is much more to be said about Jackson's knowledge argument against physicalism. But the first thing to be said is that Jackson himself has come to reject the knowledge argument. He is now convinced that physicalism must be true (2004, p. xvi):

> *On the face of it*, physicalism about the mind across the board cannot be right. [But] I now think that what is, on the face of it, true is, on reflection, false. I now think that we have no choice but to embrace some version or other of physicalism.

1.7.1 The causal argument for physicalism

David Papineau summarizes the causal argument for physicalism—which he describes as 'the canonical argument'—as follows (2002, p. 17):

> Many effects that we attribute to conscious causes have full physical causes. But it would be absurd to suppose that these effects are caused twice over. So the conscious causes must be identical to some part of those physical causes.

Following Papineau, we can set out the causal argument a little more formally.

There are three premises, of which the second is 'the completeness of physics' (2002, pp. 17–18):

(1) Conscious mental occurrences have physical effects.
(2) All physical effects are fully caused by purely *physical* prior histories.
(3) The physical effects of conscious causes aren't always overdetermined by distinct causes.

From these premises, Papineau says, materialism follows (*ibid*., p. 18)—where materialism is the thesis that conscious states are either (a) identical with physical states (in the strict sense of states of kinds studied by the physical sciences) or else (b) identical with '"physically realized functional states", or with some other kind of physically realized but not strictly physical states' (p. 15). (Papineau uses the term 'physicalism' for the stricter thesis that the first disjunct (a) is true. In previous sections, I have not distinguished between physicalism and materialism.)

It is possible to evade the causal argument by rejecting the completeness of physics (denying premise 2) and allowing, instead, that some physical occurrences have irreducibly non-physical causes. One historical view of this

kind proposed the operation of *vital forces* or special powers of living matter. Papineau (2002, Appendix) explains in some detail how developments in biochemistry and neurophysiology during the first half of the twentieth century 'made it difficult to go on maintaining that special forces operate inside living bodies' (2002, pp. 253–4). Nevertheless, Chalmers suggests that, if we were to have independent reason to reject physicalism, then we should leave open the possibility of rejecting the completeness of physics and maintaining *dualist interactionism* instead—'holding that there are causal gaps in microphysical dynamics that are filled by a causal role for distinct phenomenal properties' (2002, p. 261).

An alternative way to evade the causal argument is to accept the completeness of physics, so that all physical effects have full physical causes, but then to avoid the unwanted consequence that these effects are 'caused twice over' by denying that conscious mental occurrences have physical effects (denying premise 1). This is *epiphenomenalism* about conscious mental states, the position that Jackson (1982) adopted when he accepted the knowledge argument against physicalism.

1.7.2 Strict and relaxed versions of physicalism: identity or supervenience

Even if we accept all the premises of the causal argument as Papineau presents it, we can still evade the conclusion that conscious states are *identical* with physical states in the strict sense (physicalism, in Papineau's terminology). To see this, we need to make some distinctions within the idea of an effect being caused twice over—that is, refine the idea of overdetermination by 'distinct causes' (refining premise 3).

A man's death is overdetermined by distinct causes, or caused twice over, if he is 'simultaneously shot and struck by lightning' (2002, p. 18). That kind of causation by two independent causes is, we can agree, an unintuitive model for causation by conscious mental states. But we should also consider the case of causes that are numerically distinct but not independent. In particular, we should consider *supervenient* properties—that is, higher-level properties whose instantiation is strongly determined or necessitated by the instantiation of certain lower-level properties. As Papineau observes (*ibid.*, p. 32), it is quite natural to regard these higher-level properties as having causal powers in virtue of the causal powers of the lower-level properties on which they supervene. In short, supervenient properties may have supervenient causal powers.

Suppose we allow that some properties that supervene on physical properties might not, strictly speaking, be physical properties themselves. Then we

make room for the possibility that conscious causes might be numerically distinct from physical causes, yet without any suggestion that there would be two independent causes of the same physical effect. The causal powers of the phenomenal properties of conscious mental states would supervene on—would be determined or necessitated by (perhaps even constituted by)—the causal powers of lower-level physical properties. So it would only be in a very attenuated sense that the effects of the conscious causes would be caused twice over. It would be quite different from the case of a man being simultaneously shot and struck by lightning. It would be more like the case of a man being bruised by simultaneously bumping into both a bronze statue and the lump of bronze that constitutes the statue.[15]

Thus, in the end, the causal argument allows for a strict version of physicalism about the mind, according to which all mental properties are physical properties (defined as properties that figure in the physical sciences), and also for a more relaxed version, according to which all mental properties at least supervene on physical properties. On both identity (strict) and supervenience (relaxed) versions of physicalism, mental properties have causal powers (they are not epiphenomenal) but there is no evident threat of overdetermination by distinct and independent causes.

1.8 Jackson's rejection of the knowledge argument

The causal argument for physicalism allows for a relaxed version of physicalism—supervenience physicalism—and, as Stoljar and Nagasawa note (2004, p. 14): 'in contemporary philosophy, physicalism is usually construed in terms of what is called a supervenience thesis'. It might be tempting to assume, therefore, that the causal argument for physicalism adequately captures Jackson's reason for rejecting the knowledge argument against physicalism. However, in this section, I shall explain how Jackson's own grounds for rejecting the knowledge argument go beyond the causal argument for physicalism and how his own conception of physicalism is more demanding than supervenience physicalism.

1.8.1 Knowledge and epiphenomenalism

When he put forward the knowledge argument against physicalism, Jackson already accepted that physical effects have full physical causes (the completeness

[15] On some philosophical views (including mine), although the statue and the lump are in the same place at the same time, they are strictly speaking different objects with some different properties. Nevertheless, the man's bruise is no worse for his having bumped into both of them.

of physics) and that instantiations of non-physical properties have no causal consequences in the physical order (dualist interactionism is false). As a consequence, he accepted that phenomenal properties of experience, if they are not physical properties, are epiphenomenal. Against such a view, the causal argument for physicalism makes no headway (because premise 1 is not accepted). From the mid-1990s, Jackson came to argue (1998b, 2005a; Braddon-Mitchell and Jackson 1996) that, since the denial of physicalism involves epiphenomenalism about qualia, the knowledge argument is undermined by its own conclusion.

Jackson says that his reason for changing his mind about the knowledge argument is that (1998b/2004, p. 418): 'Our knowledge of the sensory side of psychology has a causal source.' When Mary emerges from her black-and-white room and sees something red, she undergoes a change—from not knowing what it is like to see red to knowing what it is like to see red. This is, or involves, a physical change. The physical change is caused by something and, by the completeness of physics, it has a full physical cause. If the phenomenal properties of Mary's experience of seeing a red rose or a ripe tomato are non-physical, and so epiphenomenal, then those properties of Mary's experience can play no part in the causation of Mary's coming to know what it is like to see red. Thus (Braddon-Mitchell and Jackson 1996, p. 134): 'Mary's discovery … of something important and new about what things are like is in no sense due to the properties, the qualia, whose alleged instantiation constituted the inadequacy of her previous picture of the world.'

As Jackson came to see the situation, the conclusion of the knowledge argument has the consequence that phenomenal properties are epiphenomenal, and this undermines the intuition that Mary gains new knowledge on her release as a result of experiencing for herself what it is like to see red.[16] This was enough to persuade Jackson that 'there must be a reply' to the knowledge argument (Braddon-Mitchell and Jackson 1996, p. 143), 'it *must* go wrong' (Jackson 2005a, p. 316).

1.8.2 Jackson's version of physicalism

Physicalism as Jackson conceives it is not the strict version. It is not committed to the claim that phenomenal properties are identical to properties that figure

[16] This objection against the knowledge argument was raised by Michael Watkins (1989): 'if Jackson's [1982] epiphenomenalism is correct, then we cannot even know about our own qualitative experiences' (p. 158); 'Jackson's epiphenomenalism provides us with no avenues by which we might justifiably believe that there are qualia. If epiphenomenalism is correct, then Mary, the heroine of Jackson's knowledge argument against physicalism, gains no new knowledge when she leaves her black and white room' (p. 160).

in the present or future science of physics, nor even to the claim that phenomenal properties are identical to properties that figure in the physical sciences conceived more broadly, to include physics, chemistry, biology, and neuroscience. The reason is that physicalism is 'a theory of everything in space-time' (2006, p. 231) and 'the patterns that economics, architecture, politics and very arguably psychology, pick out and theorize in terms of, include many that do not figure in the physical sciences' (*ibid*., p. 234). The properties in terms of which Jackson's physicalism is defined are not just physical properties 'in the core sense' (properties that figure in the physical sciences) but also include physical properties 'in an extended sense' (p. 233).

Jackson's version of physicalism is not the relaxed version either. The reason is this (2006, p. 243): 'A live position for dual attribute theorists is that psychological properties, while being quite distinct from physical properties, are necessitated by them.' So, supervenience physicalism characterized without some additional requirement is not properly distinguished from 'a necessitarian dual attribute view' (*ibid*.). According to Jackson, if supervenience physicalism is not to be 'a dual attribute theory in sheep's clothing' (p. 227) then the determination of supervening properties by core physical properties must be necessary *and a priori*.

Thus, Jackson proposes that physical properties in the extended sense are properties whose distribution is determined *a priori* by the distribution of physical properties in the core sense. Two simple examples may help to make this idea clearer. First, while the property of being silver and the property of being copper both figure in the science of chemistry, it is not clear that chemistry or any other physical science has a use for the disjunctive property of being either silver or copper. So the disjunctive property might not be a physical property in the core sense. But it is a physical property in the extended sense because whether something instantiates the disjunctive property is determined *a priori* by whether it is silver and whether it is copper. Second, while jewellers talk about sterling silver it is not clear that there is a science of things that are made up of 92.5% silver and 7.5% copper. So the property of being sterling silver might not be a physical property in the core sense. But it is a physical property in the extended sense because the distribution of sterling silver is determined *a priori* by the distributions of silver and of copper.

1.8.3 The case for *a priori* physicalism

Jackson's version of physicalism is *a priori physicalism* and, in fact, the notion of the *a priori* enters twice over. First, *a priori* physicalism requires that all properties should be physical properties, defined as properties that are *determined a priori* by properties that figure in the physical sciences (section 1.8.2).

Second, *a priori* physicalism requires that all the facts, particularly the psychological facts, should be *entailed a priori* by the physical facts (section 1.6.1). The two requirements are not obviously equivalent. If there were subjective conceptions of physical properties then the requirement of *a priori* entailment would not be met (there would be an explanatory gap) although the requirement of *a priori* determination of properties could still be met. In both cases, however, the *a priori* element in the account is promoted as distinguishing physicalism from 'a dual attribute theory in sheep's clothing' (2006, p. 227) or from 'some covert form of dual attribute theory of mind' (1995/2004, p. 414; see above, section 1.6.2).[17]

Jackson places his *a priori* physicalism in the tradition of Australian materialism—the materialism of J.J.C. Smart (1959) and David Armstrong (1968)—according to which 'spooky properties are rejected along with spooky substances' (2006, p. 227). He is opposed to all dual attribute theories, including even the 'necessitarian' dual attribute theory that says that phenomenal properties are distinct from, but strongly determined (necessitated) by, physical properties. The problem with dual attribute theories, Jackson says, is that 'spooky properties… would be epiphenomenal and so both idle and beyond our ken' (*ibid.*). As we saw (section 1.8.1), it was because of this problem that Jackson rejected the knowledge argument and its conclusion that the phenomenal properties of experience are 'spooky', non-physical, properties.

It may be, however, that non-physical properties need not be epiphenomenal. As Terence Horgan puts the point (1984/2004, p. 308, n. 6): 'Indeed, even if qualia are nonphysical they may not be epiphenomenal. As long as they are supervenient upon physical properties, I think it can plausibly be argued that they inherit the causal efficacy of the properties upon which they supervene.' (In section 1.7.1, we noted that Papineau (2002, p. 32) makes a similar proposal.) The possibility of non-physical, but causally potent, properties raises two potential worries about Jackson's position. First, it allows a response to Jackson's specific reason for rejecting the knowledge argument. Second, it raises the question whether there is any good objection to dual attribute theories of the necessitarian variety.

[17] Jackson says that the thesis about the *a priori* determination of physical properties is *a priori* physicalism 'understood as a doctrine in metaphysics, understood *de re* if you like' (2006, p. 229). This thesis is already sufficient to distinguish *a priori* physicalism from a necessitarian dual attribute theory. The thesis about *a priori* entailment is *a priori* physicalism understood *de dicto*. Although *a priori* physicalism understood *de dicto* seems to go beyond *a priori* physicalism understood *de re*, Jackson (2005b, p. 260) advances an argument 'that takes us from the *de re* thesis to the *de dicto* thesis'.

Jackson could respond to these worries by maintaining that his *a priori* physicalism is a better, because more austere, theory than any dual attribute view. He characterizes *a priori* physicalism as 'bare' physicalism in a passage that manifests something of W. V. O. Quine's (1953; see Jackson 2005b, p. 257) 'taste for desert landscapes' (2003/2004, pp. 425–6):

> The *bare physicalism hypothesis* … that the world is exactly as is required to make the physical account of it true in each and every detail but nothing more is true of this world in the sense that nothing that fails to follow *a priori* from the physical account is true of it … is not *ad hoc* and has all the explanatory power and simplicity we can reasonably demand.

1.9 **Physicalism and representationalism**

Jackson now rejects the conclusion of the knowledge argument. As mentioned earlier, he thinks that 'we have no choice but to embrace some version or other of physicalism' (2004, p. xvi). The specific version of physicalism that he accepts is type-A materialism—also known as *a priori* physicalism, conceptually deflationist physicalism, or bare physicalism. Consequently, he rejects the epistemological first premise of the knowledge argument. This, he now says, is where the argument goes wrong (2004, p. xvii–xviii): '[Mary] learns nothing about what her and our world is like that is not available to her in principle while in the black and white room.'

This is what Farrell would say and Dennett does say.[18] It is what Jackson needs to say; but it is not easy to defend. It certainly contrasts sharply with what he said when he first put forward the knowledge argument (1986/2004, p. 52): '[I]t is very hard to believe that [Mary's] lack of knowledge could be remedied merely by her explicitly following through enough logical consequences of her vast physical knowledge.' In defence of his new position, Jackson needs to make it plausible that, when Mary 'knows what it is like to see red', what she really knows is something that is entailed *a priori* by the totality of physical facts about the world, facts that she already knew in her room.

1.9.1 **Representationalism**

I mentioned earlier (section 1.3.1) that Armstrong interprets Farrell's claim about experience being featureless as an anticipation of the representationalist view that the phenomenal properties of an experience are determined by its representational properties—that is, by how it represents the world as being. Similarly, Stoljar and Nagasawa (2004, p. 25, n. 11) see Farrell's idea of featureless

[18] See Dennett (1991, pp. 398–406, reprinted in Ludlow *et al.* 2004, pp. 59–68; 2005, Chapter 5, 'What RoboMary knows').

experience as related to the doctrine of the transparency or diaphanousness of experience, a doctrine that several contemporary philosophers take to stand in a close relationship to representationalism.[19] In any case, it is to representationalism that Jackson turns for his account of what Mary really knows when she knows what it is like to see red (2003/2004, p. 430): 'we have to understand the qualities of experience in terms of intensional [representational] properties'.

The starting point for representationalism is that 'experience is *essentially* representational … it is impossible to have a perceptual experience without thereby being in a state that represents that things are thus and so in the world' (Jackson 2007, p. 57). As a claim about perceptual experience, such as Mary's experience of a red rose or a ripe tomato, this is highly plausible. It is of the nature of perception to represent how things are in the world. The claim that *all* experiences, including bodily sensations, are representational is less compelling but even strong opponents of representationalism may be prepared to grant it. Thus, for example, Ned Block says, 'I think that sensations—almost always—perhaps even always—have representational content' (2003, p. 165).

Going beyond this starting point, representationalism is usually formulated as a supervenience thesis. As between two conscious mental states, such as two perceptual experiences, there can be no difference in phenomenal properties without a difference in representational properties. The phenomenal character of a conscious mental state is determined by its representational content (Byrne 2001). Representationalism is an unclear thesis to the extent that the notion of representation itself is not well specified. It is also a controversial thesis. But the attraction of representationalism for a type-A materialist is that it promises physical-functional conceptions of types of experience.

In philosophy of mind over the last quarter-century or so, the topics of consciousness and representation have mainly been considered somewhat separately. As a result, even those who think that consciousness defies scientific explanation are apt to be confident that representation can be analysed in 'naturalistic', physical-functional, terms. It is against this background that Jackson says (2003/2004, p. 432):

> The project of finding an analysis of representation is not an easy one—to put it mildly. But … the answers that have been, or are likely to be, canvassed are all answers

[19] It is difficult to spell out a compelling argument from the transparency or diaphanousness of experience to representationalism (Stoljar 2004; see also Burge 2003, pp. 405–7). Tye (2000, p. 45) says: 'I believe that experience is transparent. I also believe that its transparency is a very powerful motivation for the representationalist view. I concede, however, that the appeal to transparency has not been well understood.' Jackson (2007, p. 57) says: 'I conclude that the famous diaphanousness or transparency of experience is not *per se* the basis of an argument for representationalism.'

that would allow the fact of representation to follow *a priori* from the physical account of what our world is like.

1.9.2 Representationalism and phenomenology

Let us agree, for the sake of the argument, that the representational facts are entailed *a priori* by the physical facts, just as the water facts are entailed *a priori* by the H_2O facts. This is not yet sufficient for a physicalist account of what Mary really knows when she 'knows what it is like to see red'. For what Mary knows entails that *there is something that it is like* to see red; seeing red is a conscious mental state. But, even according to representationalism, it is *not* the case that the representational content of an experience 'suffices to make any state that has it conscious' (Seager and Bourget 2007, p. 263). Or, as Alex Byrne puts it (2001, p. 234): 'Intentionalism [representationalism] isn't much of a theory of consciousness.'

Representationalism says that the phenomenal character of a conscious mental state is determined by its representational content. But it does not say that the representational properties of a mental state determine that it is a *conscious* mental state. There can be representation without consciousness. So something needs to be added to representationalism if it is to provide an account of what Mary knows about the experience of seeing red. Jackson himself says that the nature of an experience, including the fact that it is a conscious mental state, is determined by 'the [representational] content of [the] experience *plus* the fact that the experience represents the content as obtaining *in the way distinctive of perceptual representation*' (2007, p. 58; also see 2005a, p. 323).

He goes on to list five features that are putatively distinctive of perceptual representation. The content of perceptual representation is rich, and inextricably rich; the representation is immediate; there is a causal element in its content; and perceptual experience plays a distinctive functional role in respect of belief. If a state has representational content with these five features then, Jackson says, 'we get the phenomenology for free' (2003/2004, p. 438). What is most important about these five features is that they can, let us suppose, be explicated in physical-functional terms.

The story of Mary generates the intuition that, on her release, Mary learns something new about the experiences of people who looked at red roses and ripe tomatoes while she was in her room. As required by type-A materialism, Jackson rejects this intuition. He says that what Mary really knows is that the people were in physical states with a particular representational property (roughly, representing something as being red) and meeting five further conditions. Since both representation and the further conditions can be

explicated in physical-functional terms, this knowledge is not new but was, in principle, already available to Mary while she was in her room.

1.10 The epistemological intuition and the ability hypothesis

Type-A materialism is counter-intuitive. Accepting it commits Jackson to rejecting the epistemological intuition that Mary learns something new on her release. He needs to develop a critical agenda supporting that rejection. Like Farrell and Dennett, Jackson needs to undermine the idea that there is more to know about human experience than is entailed *a priori* by the totality of physical facts.

1.10.1 Representationalism and the epistemological intuition

Jackson stresses that representationalism highlights the distinction between a *representational* property and an *instantiated* property (2003, 2005a, 2007). Representationalism thus provides a reason to say that 'there is no such property' (2003/2004, p. 430) as the 'redness' of the experience of seeing a rose. Redness is not a property that experiences of roses *instantiate*; it is the property that experiences *represent* roses as instantiating.

Experiences do, of course, instantiate the property of representing things as being red. But this representational property of experiences is not a new property that was unknown to Mary before her release. It is a physical-functional property that Mary knew about (or could have known about) in her black-and-white room. What is new after her release is that Mary now has an experience instantiating this property (and meeting five further conditions). We must take care not to mistake a new instantiation of a representational, and therefore physical, property of experiences for the instantiation of a new, and therefore non-physical, property of experiences. A new instantiation of a property is not the instantiation of a new property.

This is an important point, but it is not clear that it undermines the intuition that supports the epistemological first premise of the knowledge argument (Alter 2007). Jackson says (2007, p. 61): 'The challenge from the knowledge argument is the intuition that the "red" of seeing red is a new sort of property.' But, on the face of it, the intuition that Jackson needs to undermine—the intuition that drives the knowledge argument—is not an explicitly metaphysical intuition that, on her release, Mary learns about a new property of experiences. It is the epistemological intuition that Mary gains new knowledge—that she comes to know a fact that is not entailed *a priori* by the totality of physical facts that she already knew in her room.

It is, of course, part of Jackson's overall position that new knowledge would have to be knowledge about new properties. That is what the second premise of the original knowledge argument says: if physicalism is true (no new properties) then the psychophysical condition is *a priori* (no new knowledge). But that connection between epistemology and metaphysics is not provided by representationalism about perceptual experiences. It depends on the assumption that type-B materialism can be rejected (section 1.6.3).

Representationalism may be developed in the service of physicalism about conscious mental states. It certainly plays a major role in contemporary philosophy of consciousness. But representationalism does not favour Jackson's conceptually deflationist physicalism (type-A materialism) over conceptually inflationist physicalism (type-B materialism). Tye (1995, 2000) develops a version of representationalism that is very similar to Jackson's.[20] Yet Tye maintains that, on her release, Mary gains new subjective conceptions of physical—specifically, representational—properties, deploys those conceptions in propositional thinking, and achieves new propositional knowledge.

1.10.2 The ability hypothesis

We have just seen that representationalism does not provide any independent motivation for rejecting the epistemological intuition that Mary learns something new on her release. But Jackson's physicalist account of conscious experience goes beyond representationalism.

According to representationalism, the properties of Mary's experience when she sees a red rose for the first time are physical properties. Specifically, they are properties of having such-and-such representational content and meeting further physical-functional conditions. They are not new properties but properties that Mary was already in a position to know about in her black-and-white room. What is *new* is that Mary now has an experience that *instantiates* those properties. Jackson describes Mary's situation as follows (2003/2004, p. 439):

> [S]he is in a new kind of representational state, different from those she was in before. And what is it to know what it is like to be in that kind of state? Presumably, it is to be able to recognize, remember, and imagine the state. … We have ended up agreeing with Laurence Nemirow and David Lewis on what happens to Mary on her release.

[20] According to Tye's PANIC theory, phenomenal properties are determined by (indeed, are identical with) properties of having Intentional Content that meets three conditions: it is Poised (poised to have a direct impact on beliefs and desires), Abstract (does not involve particular objects), and Nonconceptual (in order for a state to have this kind of content it is not necessary for the subject of the state to be able to conceptualise the content). Thus, in Tye's account, P+A+N plays the role that the five features play in Jackson's account.

Here, Jackson goes beyond the basic claim of his response to the knowledge argument, namely, that Mary does not gain new propositional knowledge on her release. He concedes something to the epistemological intuition. He says that Mary does gain something new and he allows that this might be described as new 'knowing what it is like'. But following the *ability hypothesis response* to the knowledge argument proposed by Lewis (1988) and Nemirow (1980, 2007), he says that what Mary gains is not new propositional knowledge but only new abilities.

Papineau comments on the ability hypothesis (2002, pp. 59–60):

> Some philosophers are happy to accept that Mary acquires new powers of imaginative re-creation 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 of the 'ability hypothesis'.

Jackson proposes to join these 'sophisticated' conceptually deflationist physicalists and, as Papineau says, their position—Mary gains new 'know how' or new abilities—certainly seems preferable to 'outright denial' that Mary comes to know anything new at all. Subtle type-A materialism seems preferable to the more straightforward type-A materialism of Farrell and Dennett. But, Papineau continues (*ibid*., p. 61):

> Even so, the ability hypothesis does not really do justice to the change in Mary. If we look closely at Mary's new abilities, we will see that they are inseparable from her power to think certain new kinds of thoughts.

The abilities to which Jackson appeals in his description of Mary's new 'knowing what it is like' are the same abilities to which Tye (2000) appeals in describing a subject's possession of a subjective conception of a relatively coarse-grained type of experience, such as seeing red (section 1.2.2). The advocate of the ability hypothesis needs to say why having these abilities is not sufficient for possession of a subjective conception that can be deployed in propositional thinking about a type of experience.

In earlier writings, Jackson himself resists the ability hypothesis response to the knowledge argument. He argues that Mary's new abilities (to remember, imagine, recognize) are associated with a new cognitive capacity to engage in new propositional thinking and that Mary acquires 'factual knowledge about the experiences of others' (1986/2004, p. 55). For example, he says (*ibid*., p. 54):

> Now it is certainly true that Mary will acquire abilities of various kinds after her release. She will, for instance, be able to imagine what seeing red is like, be able to remember what it is like … But is it plausible that that is *all* she will acquire? … On her release she sees a ripe tomato in normal conditions, and so has a sensation of red. Her first reaction is to say that she now knows more about the kind of experience others have when looking at ripe tomatoes.

Jackson now embraces the ability hypothesis but it seems to me that his earlier line of argument is still plausible.

Even if the ability hypothesis can be defended against these objections, its role is only to provide a more nuanced version of type-A materialism, reducing the counterintuitive impact of denying that Mary gains new propositional knowledge on her release. It does not provide any strong, independent reasons in favour of type-A materialism and against type-B materialism.[21] Thus, neither representationalism nor the ability hypothesis offers materials for the critical agenda that any type-A materialist needs to develop. They do nothing to undermine the idea that there is more to know about human experience than is entailed *a priori* by the totality of physical facts.

1.11 Physical properties in new guises?

If physicalism is true then phenomenal properties are physical properties. Specifically, if Jackson's *a priori* physicalism is true then phenomenal properties are physical properties either in the core sense (properties that figure in the physical sciences) or in the extended sense (properties whose distribution is determined *a priori* by the distribution of physical properties in the core sense). If Jackson's or Tye's representationalism is true then we can say more about phenomenal properties: they are representational properties. The leading idea of type-B materialism is that Nagelian subjective conceptions of types of experience are conceptions *of* physical properties. The explanatory gap is consistent with physicalism; there is habitable space between epistemological leaving out and metaphysical leaving out; there can be new knowledge that is not knowledge of new properties.

1.11.1 The 'old fact, new guise' response to the knowledge argument

Type-B materialism is conceptually inflationist physicalism. It involves a duality of objective and subjective conceptions without a metaphysical duality of physical and non-physical properties. The type-B materialist's response to the knowledge argument is to accept the epistemological first premise but deny that this leads to the metaphysical conclusion that physicalism is false. When Mary is released from her black-and-white room, her new knowledge involves a new subjective conception of a fact that she already knew under an old

[21] Yuri Cath (in press) argues that the ability hypothesis leads ultimately to the idea of new conceptions, and so turns out to be a version of the type-B materialist's 'old fact, new guise' response to the knowledge argument (see below, section 1.11).

objective conception. As Levine says (1993, p. 125): 'the case of Mary typifies the phenomenon of there being several distinguishable ways to gain epistemic access to the same fact'.

This is a common response to the knowledge argument (e.g. Horgan 1984), often known as the 'old fact, new guise' response (or the 'old fact, new mode' response, to suggest Gottlob Frege's (1892) notion of mode of presentation). Jackson rejects it, both early and late. When he first put forward the knowledge argument against physicalism, he already rejected the suggestion that the argument depends on 'the intensionality of knowledge' (1986/2004, p. 52).[22] He now accepts physicalism and is convinced that the knowledge argument '*must* go wrong'. But he still rejects the suggestion that the argument goes wrong in neglecting new conceptions, guises, or modes of presentation (Braddon-Mitchell and Jackson 2007, p. 137): 'This is the explanation of Mary's ignorance that is available to dual attribute theorists, not the explanation available to physicalists.'

There can certainly be multiple conceptions of the same physical property. But Jackson maintains that explaining Mary's new knowledge by appeal to new conceptions is incompatible with physicalism. In order to understand why, it is useful to recall the example of water and H_2O (section 1.6.2). On Jackson's view, it is a matter of conceptual analysis that water is the stuff that fills the water role. Knowledge that water is H_2O can only be arrived at *a posteriori*. But the fact that water is H_2O is entailed *a priori* by the fact that H_2O fills the water role. Now suppose that some physical stuff, S, fills two roles, R_1 and R_2, in the physical order. Then we can have two conceptions of S. We can think of S as the stuff that fills role R_1 or as the stuff that fills role R_2. It may very well be that examining these conceptions themselves will not tell us that they are two conceptions of the same physical stuff. It is likely that this knowledge can only be arrived at *a posteriori*. But if Mary, in her black-and-white room, knows the full story about the physical order, then she is already in a position to know that the stuff that fills role R_1 also fills role R_2. This kind of example of multiple conceptions is available to a physicalist, but it does not provide a model for Mary's gaining new knowledge on her release.

The situation would be different if S were to instantiate a non-physical property, N. Then we could have a third conception of S. We could think of S as the stuff that instantiates N. Even if Mary knew all there is to know about

[22] The intensionality of knowledge is illustrated by the fact that 'Nigel knows that Hesperus is a planet' may be true while 'Nigel knows that Phosphorus is a planet' is false even though Hesperus = Phosphorus.

the physical order, she might not be in a position to know that the stuff that fills role R_1 and role R_2 also instantiates N. In her black-and-white room, she might know nothing at all of property N. But, while this kind of example of multiple conceptions would provide a model for Mary's gaining new knowledge on her release, it is obviously not available to a physicalist (Jackson 2005b, p. 262).

1.11.2 Descriptive and non-descriptive conceptions

According to the conceptually inflationist physicalist (type-B materialist), Mary gains new knowledge on her release because she gains new subjective conceptions of physical properties. The problem for type-B materialism is that we have not been able to find a model for Mary's new conceptions that is consistent with physicalism.[23]

It is plausible that the source of this problem lies in the fact that we have considered only conceptions that pick out a kind of physical stuff *by description*. These are conceptions of the form 'the physical stuff that has property F', where the property in question might be a physical property or a non-physical property. Descriptive conceptions in which the descriptive property is physical or physical-functional (such as 'the physical stuff that *fills the water role*') are already available to Mary while she is in her room. So it may seem that new, distinctively subjective conceptions must be descriptive conceptions in which the descriptive property is non-physical (such as 'the physical stuff that has *property N*'). Thus, if all conceptions are descriptive then Nagel's duality of objective and subjective conceptions requires a metaphysical duality of physical and non-physical properties right from the outset. Since type-B materialism proposes a duality of conceptions without a duality of properties, it requires that subjective conceptions—including the conceptions that become available to Mary only on her release—are *not* descriptive conceptions.

A partial analogy for the distinction between objective conceptions and non-descriptive subjective conceptions is provided by the distinction between two kinds of conception of locations in space. One kind of conception specifies locations in terms of distances and directions from an objective point of origin. A location, L, might be specified as being 25 miles north-west from Carfax. Deploying that conception of L in thought, I might achieve propositional

[23] See Jackson (2005a, p. 318): '[T]he guises … must all be consistent with physicalism if physicalism is true…. But then, it seems, Mary could know about their applicability when inside the room.'

knowledge (by looking at a map or reading a book, for example) that there is water at L—perhaps the book says that there is a pond with ducks. A different kind of conception of a location is made available to me when I am *at* that location. Without knowing how far I am from Carfax or in which direction, I might arrive at a location and decide to explore a little. What is going on here? I notice sheep grazing on the other side of a stone wall, some farm buildings further back, a tractor and, in the distance, trees. Then I see a pond with ducks. So, there is water here.

If I am, in fact, 25 miles north-west from Carfax then this is an example of a new instantiation of an old spatial property: I myself am at location L. In virtue of my new location, I gain new abilities: I can feed the ducks and, just by bending down, I can put my hand in the pond, I can see (and may later remember) things that I have never seen before. But I do not only gain new abilities. I also gain a new indexical or egocentric conception of a location that I already knew about under a different, map-based, 'distance and direction from origin' or allocentric conception. I have a new cognitive capacity: I can think of location L as 'here'. Deploying the new conception in propositional thinking, I achieve new propositional knowledge that (as I put it) 'there is water here'.

We should not rush from this partial analogy to the idea that subjective conceptions of types of experience are indexical conceptions, like the context-dependent conceptions of locations, times, and people expressed by 'here', 'now', and 'I'. In fact, there are reasons to reject the proposal that subjective conceptions are indexical conceptions (Tye 1999; Papineau 2007). One disanalogy is that at least some subjective conceptions can be deployed in thought by a subject who is not concurrently having an experience of the type in question. They seem to function like recognitional, rather than indexical, conceptions (Loar 1997). But, in the face of Jackson's objection to the 'old fact, new guise' response, even the partial analogy offers some encouragement to the conceptually inflationist physicalist.

1.12 Phenomenal concepts and physicalism

Non-descriptive subjective conceptions of phenomenal types of experience are often called *phenomenal concepts*. Papineau introduces the idea with three main points. First, he says (2002, p. 48): '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*'. Second, he stresses that 'as a materialist, I hold that even phenomenal concepts refer to material *properties*' (*ibid.*). Third, he insists that the advocate of phenomenal concepts must avoid the 'poisoned chalice' (p. 86)

of considering phenomenal concepts as descriptive concepts. Phenomenal concepts refer 'directly, and not via some description' (p. 97).[24]

In an earlier and seminal paper, Brian Loar says (1997/2004, p. 219):

> On a natural view of ourselves, we introspectively discriminate our own experiences and thereby form conceptions of their qualities, both salient and subtle …. What we apparently discern are ways experiences differ and resemble each other with respect to *what it is like to have them*. Following common usage, I will call these experiential resemblances *phenomenal qualities*, and the conceptions we have of them, *phenomenal concepts*. Phenomenal concepts are formed 'from one's own case'.

Loar goes on to highlight the distinction between concepts and properties and to point to the possibility of accepting a duality of concepts or conceptions without a duality of properties (1997/2004, pp. 220–1):

> It is my view that we can have it both ways. We may take the phenomenological intuition at face value, accepting introspective concepts and their conceptual irreducibility, and at the same time take phenomenal qualities to be identical with physical-functional properties of the sort envisaged by contemporary brain science. As I see it, there is no persuasive philosophically articulated argument to the contrary.

1.12.1 A limitation on the promise of phenomenal concepts

Phenomenal concepts provide a model for subjective conceptions of types of experience—including the new conceptions that Mary gains on her release—and the model holds some promise of being consistent with physicalism. First, according to type-B materialism, subjective conceptions are conceptions *of* physical properties. Second, a phenomenal concept of a type of experience is a non-descriptive concept. It is a recognitional concept that a thinking subject possesses in virtue of having an experience of the type in question. Deploying a phenomenal concept in thought is not a matter of thinking of a physical property as the property that has such-and-such higher-order property (that is, such-and-such property of properties). So the type-B materialist need not face an objection along the lines that phenomenal concepts can only account for new knowledge if they involve *non-physical* higher-order properties (section 1.11.2). Nevertheless, the promise of phenomenal concepts is limited in an important way.

According to physicalism, conscious mental states are physical states and the phenomenal properties of conscious mental states are physical properties.

[24] In *Thinking About Consciousness* (2002), Papineau defends a 'quotational' or 'quotational-indexical' model of phenomenal concepts. More recently (2007), he acknowledges that this model faces some objections and he adopts a different view of phenomenal concepts as cases of, or at least as similar to, perceptual concepts—something like 'stored sensory templates' (2007, p. 114). This change leaves intact the three points in the main text.

In general, instantiating a physical property is not sufficient for gaining a conception of that property but, according to type-B materialism, the phenomenal properties of conscious mental states are special in this respect. Consider a subject who is, in general, able to form concepts and deploy them in thought. By being in a conscious mental state—having an experience of a particular type—such a subject can gain conceptions of certain physical properties of that state, namely, the phenomenal properties of that type of experience. These conceptions are direct, non-descriptive, subjective, phenomenal concepts and, intuitively, the subject gains a phenomenal concept of a physical property in such cases only because there is something that it is like to instantiate that property.

Now, recall Nagel's remark (1974/1997, p. 524):

> If mental processes are indeed physical processes, then there is something that it is like, intrinsically, to undergo certain physical processes. What it is for such a thing to be the case remains a mystery.

According to Nagel, if there is something that it is like to instantiate certain physical properties then we have no answer to the question *why* this is so.

We have just said that, intuitively, if we gain phenomenal concepts of certain physical properties by instantiating them then there must be something that it is like to instantiate those properties. Consequently, it seems, if we gain phenomenal concepts of certain physical properties by instantiating them then, ultimately, we have no answer to the question why that is so. Possessing phenomenal concepts of physical properties does not have a fully satisfying explanation in physical terms.[25]

1.12.3 An argument against type-B materialism?

This limitation on the promise of phenomenal concepts may offer the prospect of an argument against the type-B materialist's claim that phenomenal concepts are direct, non-descriptive concepts of physical properties.

A subject who has a phenomenal concept of a type of experience meets the requirement of *knowing which* type of experience is in question (section 1.2.2). The subject knows what that type of experience is like (in one use of 'knowing what it is like') in virtue of being, or having been, the subject of an experience of that type. But, a subject who knows which type of experience is in question need not think of that type of experience as the property with such-and-such physical-functional specification nor, indeed, as being a physical property at all.

[25] For discussion of what can reasonably be demanded of the phenomenal concept response to the knowledge argument, see Chalmers (2007), Levine (2007), Papineau (2007).

Furthermore, if Nagel is right then a subject who knows what a type of experience is like has no answer to the question why this is what it is like, or why there is anything at all that it is like, to instantiate a property whose nature is physical.

We might begin to wonder whether a subject can really possess a direct and non-descriptive concept *of a physical property*, and meet the requirement of knowing which physical property is in question, just in virtue of knowing what a particular type of experience is like. A subject who knows what a type of experience is like has a phenomenal concept. But we might wonder whether *thinking about a physical property* by deploying a phenomenal concept must be *indirect*, with the phenomenal concept embedded in a descriptive concept along the lines of: 'the physical property that it is *like this* to instantiate'.[26]

Developing, and then responding to, these inchoate concerns would require work on the metaphysics of properties—their individuation and their natures—and work on the 'knowing which' requirement that would inevitably lead into theories of reference in philosophy of language and thought. It is not obvious in advance what the outcome of this work would be. But suppose, for a moment, that it were to uncover a good argument for the claim that, if there are distinctively subjective phenomenal concepts of phenomenal properties, then these phenomenal properties are not identical with physical properties (they neither are, nor are determined *a priori* by, properties that figure in the physical sciences).

This would be an important argument. First, it would show that type-B materialism can be rejected—that a duality of objective and subjective conceptions requires a duality of physical and non-physical properties—and it would show this without simply relying on an assumption that all conceptions are descriptive (section 1.11.2). Second, by showing that type-B materialism can be rejected, the argument would provide the needed motivation for the second premise of the original knowledge argument (section 1.6.2) and—what comes to the same thing—it would license the transition from the limited conclusion of the simplified knowledge argument, that type-A materialism is false, to the more sweeping conclusion of the original knowledge argument, that physicalism is false (section 1.6.4).

Third, the argument would tie together the two requirements of Jackson's *a priori* physicalism (section 1.8.3). The first requirement is that all properties should be physical properties, defined as properties that are *determined*

[26] Chalmers (1999) and Horgan and Tienson (2001) argue against the claim that direct phenomenal concepts are concepts of physical-functional properties. Also recall McGinn's comment that 'if we know the essence of consciousness by means of acquaintance, then we can just see that consciousness is not reducible to neural or functional processes' (2004, p. 9).

a priori by properties that figure in the physical sciences. The second requirement is that all the facts should be *entailed a priori* by the physical facts. Suppose that the first requirement is met, so that physical properties are all the properties there are. It would follow from the envisaged argument that there are no distinctively subjective conceptions of any properties. But if all properties are physical properties and there are no subjective conceptions, then there is no impediment to *a priori* entailment of all the facts by the physical facts. So the second requirement would also be met (cf. Jackson 2005b, p. 260).

Fourth, the argument would figure as an item on the critical agenda that any type-A materialist needs to develop. Jackson needs to undermine the intuition that there is more to know about human experience than is entailed *a priori* by the totality of physical facts. While representationalism is consistent with physicalism, it does not reveal any error or confusion in the epistemological intuition that drives the knowledge argument (section 1.10.1). Nor does the ability hypothesis provide strong, independent reasons in favour of type-A materialism (section 1.10.2). By showing that type-B materialism can be rejected, the argument would reveal that, even if type-A materialism involves some cost to intuition, it is the only alternative to dualism.

1.12.4 Options for physicalism

Following Chalmers, we have divided physicalist approaches to the philosophy of consciousness into two varieties, type-A materialism (also known as conceptually deflationist physicalism) and type-B materialism (also known as conceptually inflationist physicalism). We have just considered, in a speculative way, a possible line of argument against type-B materialism. If there were to be a good argument of the envisaged kind then the options would seem to be severely limited. A physicalist, having rejected the dualist options of interactionism and epiphenomenalism, would seem bound to embrace the counterintuitive commitments of type-A materialism.

In fact, this is not quite right. At the beginning of this chapter, when I first contrasted dualism and physicalism (section 1.1.1), I said that, according to physicalism, phenomenal properties are either identical with physical properties or else strongly determined (necessitated) by physical properties. I also said that the causal argument for physicalism allows for both a strict identity version and a relaxed supervenience version of physicalism (section 1.7.2). In recent sections, however, I have adopted Jackson's terminology. His version of physicalism says that all properties are physical properties. He allows that physical properties include properties that do not themselves figure in the physical sciences but are determined *a priori* by properties that do figure there. He does *not* allow that properties that are determined or necessitated only *a posteriori*

by physical properties are themselves physical. Some varieties of supervenience *physicalism* are now classified as varieties of *dualism* and, specifically, as necessitarian dual attribute theories.

This means that we need to reconsider the hypothetical situation if there were to be a good argument against the type-B materialist's claim that phenomenal concepts are direct, non-descriptive concepts of physical properties. If the dualist options of interactionism and epiphenomenalism were rejected there would still be *two* options available, not only type-A materialism, but also the necessitarian dual attribute view. Theorists who describe this view as a variety of supervenience physicalism, rather than dualism, will regard it as conceptually inflationist, rather than deflationist, physicalism. As a consequence, it will be grouped with type-B materialism. But it will, apparently, be left untouched by the line of argument against type-B materialism that we considered in section 1.12.3. According to the necessitarian dual attribute view, phenomenal concepts are not concepts of physical properties, but concepts of distinct phenomenal properties that supervene on physical properties.

The costs and benefits of the variety of supervenience physicalism also known as the necessitarian dual attribute view are not, of course, affected by a terminological decision between 'physicalism' and 'dualism'. In this chapter, we have seen only one argument against this option and that was an Ockhamist[27] argument in favour of the austerity of 'bare physicalism' (section 1.8.3). The benefits of austerity would have to be weighed against the costs to intuition of type-A materialism.

The less austere, but otherwise more intuitive, option is favoured by Edmund Rolls (this volume), who leaves it as an open question whether it is best described as 'physicalism'. According to his higher-order syntactic thought (HOST) theory of consciousness, conscious mental states are physical states of a system with a particular computational nature. The computational properties of the state necessitate phenomenal properties *a posteriori* (p. 154; some emphases added):

> [T]he present approach suggests that it *just is* a property of HOST computational processing with the representations grounded in the world that it feels like something. There is to some extent *an element of mystery* about why it feels like something, *why it is phenomenal* … In terms of the physicalist debate, an important aspect of my proposal is that it is a *necessary* property of this type of (HOST) processing that it feels like something… and given this view, then it is *up to one to decide whether this view is consistent with one's particular view of physicalism or not*.

[27] The fourteenth-century philosopher, William of Ockham, is credited with a law of parsimony: 'Entities should not be multiplied beyond necessity.'

1.13 **Conclusion**

At the first choice point in the philosophy of consciousness, some philosophers deny that there is an explanatory gap and accept type-A materialism. We have seen that Jackson joins Farrell and Dennett in this group, rejecting the intuition that there is more to know about human experience than what is entailed *a priori* by a battery of physical fact and theory that can be grasped by Mary in her black-and-white room, or by a sufficiently intelligent Martian or bat. Philosophers who, at the first choice point, accept that there is an explanatory gap proceed to a second choice point. There, some opt for dualism, others for type-B materialism.

Jackson's knowledge argument and Chalmers's conceivability argument are arguments for dualism. If these arguments are correct then the phenomenal properties of experience are not physical properties. Philosophers who accept this conclusion—Jackson (at an earlier stage when he accepted the knowledge argument) and Chalmers (still)—face a further choice about the causal relationship between the phenomenal and the physical. One option is to accept dualist interactionism at the cost of rejecting the completeness of physics. Another is to accept epiphenomenalism at the cost of rendering phenomenal properties 'idle and beyond our ken', as Jackson (2006, p. 227) now puts it. These are not especially attractive views but Chalmers (2002) argues that these options, and others, should be taken seriously 'if we have independent reason to think that consciousness is irreducible' (2002, p. 263).[28]

Chalmers also commends a view, *Russellian* or *type-F monism* (Russell 1927), on which the most fundamental properties of the physical world are both protophysical and protophenomenal—the physical and the phenomenal are variations on a common theme (2002, p. 265–6): 'One could give the view in its most general form the name *panprotopsychism*, with either protophenomenal or phenomenal properties underlying all of physical reality.' This view is speculative and exotic, but Chalmers suggests that 'it may ultimately provide the best integration of the physical and the phenomenal within the natural world' (p. 267; see also Stoljar 2006).

According to type-B materialism, we can accept that consciousness is conceptually irreducible but reject dualism. This is an attractive option that is adopted by many contemporary philosophers of consciousness—probably the majority—including Block, Levine, Loar, Papineau and Tye. If type-B materialism can be defended then arguments for dualism are undermined and some

[28] Chalmers (2002) refers to dualist interactionism as *type-D dualism* and to epiphenomenalism as *type-E dualism*.

of the motivation for more exotic views, such as Russellian monism, is removed. Indeed, the knowledge argument against physicalism and in favour of dualism seems to rest on the assumption that type-B materialism is not a real option for the physicalist.

There are, however, arguments against type-B materialism—against the idea that we can have a duality of objective and subjective concepts, and an explanatory gap, without a duality of physical and non-physical properties. Some of these arguments seem to depend on the assumption that all concepts are descriptive and the dominant form of type-B materialism therefore appeals to direct, non-descriptive, subjective, phenomenal concepts of physical properties. But there are also arguments against this form of the view.

It may very well be that none of these arguments is, in the end, compelling and that Loar (1997) will turn out to be right in saying that we can 'have it both ways': irreducibly subjective phenomenal concepts are nevertheless concepts of physical properties of the kinds that figure in neuroscience. On the other hand, there may be a good argument against phenomenal concepts of physical properties and friends of type-B materialist may have to consider shifting to the necessitarian dual attribute view that phenomenal concepts are concepts of non-physical phenomenal properties that are determined or necessitated—but not *a priori*—by physical properties. Some philosophers may object to the departure from ontological austerity (Jackson 2003) and others may have concerns about a primitive relation of *a posteriori* necessitation between properties (strong necessities; see Chalmers 1996, 1999, 2002). But if, at the first choice point, there are good reasons to accept that there is an explanatory gap then, at the second choice point, the necessitarian dual attribute view should be taken at least as seriously as Russellian monism, dualist interactionism, or epiphenomenalism.

Acknowledgements

I am grateful to Tim Bayne, Ned Block, Tyler Burge, Alex Byrne, David Chalmers, Frank Jackson, David Papineau, Edmund Rolls, Nick Shea, Daniel Stoljor, and Larry Weiskrantz for comments and conversations.

References

Alter, T. (2007). Does representationalism undermine the knowledge argument? In Alter, T. and Walter, S. (eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, pp. 65–76. Oxford: Oxford University Press.

Armstrong, D.M. (1968). *A Materialist Theory of the Mind*. London: Routledge & Kegan Paul.

Armstrong, D.M. (1996). Qualia ain't in the head. Review of *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*, by Michael Tye. *Psyche* 2(31); http://psyche.cs.monash.edu.au/v2/psyche-2-31-armstrong.html.

Block, N. (1978). Troubles with functionalism. In Wade Savage, C. (ed.) *Perception and Cognition: Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science*, Volume 9, pp. 261–325. Minneapolis: University of Minnesota Press. Reprinted in Block, N. (ed.) *Readings in the Philosophy of Psychology*, Volume 1, pp. 268–306. Cambridge, MA: Harvard University Press, 1980.

Block, N. (2002). The harder problem of consciousness. *Journal of Philosophy* **99**, 391–425.

Block, N. (2003). Mental paint. In Hahn, M. and Ramberg, B. (eds) *Reflections and Replies: Essays on the Philosophy of Tyler Burge,* pp. 165–200. Cambridge, MA: MIT Press.

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences* **9**, 46–52.

Block, N. Flanagan, O., and Güzeldere, G. (eds) (1997). *The Nature of Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press.

Braddon-Mitchell, D. and Jackson, F.C. (1996). *Philosophy of Mind and Cognition: An Introduction*. Oxford: Blackwell.

Braddon-Mitchell, D. and Jackson, F.C. (2007). *Philosophy of Mind and Cognition: An Introduction*, 2nd edn. Oxford: Blackwell.

Burge, T. (1997). Two kinds of consciousness. In Block, N., Flanagan, O., and Güzeldere, G. (eds) *The Nature of Consciousness: Philosophical Debates*, pp. 427–434. Cambridge, MA: MIT Press.

Burge, T. (2003). Qualia and intentional content: Reply to Block. In Hahn, M. and Ramberg, B. (eds) *Reflections and Replies: Essays on the Philosophy of Tyler Burge*, pp. 405–616. Cambridge, MA: MIT Press.

Burge, T. (2007). Reflections on two kinds of consciousness. In *Foundations of Mind: Essays by Tyler Burge*, Volume 2, pp. 392–419. Oxford: Oxford University Press.

Byrne, A. (2001). Intentionalism defended. *Philosophical Review* **110**, 199–240.

Byrne, A. (2006). Review of *There's Something About Mary*, by Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar. *Notre Dame Philosophical Reviews* (20.1.2006) http://ndpr.nd.edu/review.cfm?id=5561.

Cath, Y. (in press). The ability hypothesis and the new knowledge-how. *Noûs*.

Chalmers, D.J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, **2**, 200–19. Reprinted as 'The hard problem of consciousness' (pp. 225–235) and 'Naturalistic dualism' (pp. 359–368) in Velmans, M. and Schneider, S. (eds) *The Blackwell Companion to Consciousness*. Oxford: Blackwell, 2007.

Chalmers, D.J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chalmers, D.J. (1999). Materialism and the metaphysics of modality. *Philosophy and Phenomenological Research* **59**, 473–496.

Chalmers, D.J. (2000). What is a neural correlate of consciousness? In Metzinger, T. (ed.) *Neural Correlates of Consciousness: Empirical and Conceptual Issues*, pp. 17–39. Cambridge, MA: MIT Press.

Chalmers, D.J. (2002). Consciousness and its place in nature. In Chalmers, D.J. (ed.) *Philosophy of Mind: Classical and Contemporary Readings*, pp. 247–272. Oxford: Oxford University Press.

Chalmers, D.J. (2007). Phenomenal concepts and the explanatory gap. In Alter, T. and Walter, S. (eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, pp. 167–194. Oxford: Oxford University Press.

Davies, M. and Humberstone, I.L. (1980). Two notions of necessity. *Philosophical Studies* **38**, 1–30.

Dennett, D.C. (1988). Quining qualia. In Marcel, A.J. and Bisiach, E. (eds) *Consciousness in Contemporary Science*, pp. 42–47. Oxford: Oxford University Press.

Dennett, D.C. (1991). *Consciousness Explained*. Boston: Little, Brown.

Dennett, D.C. (2005). *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. Cambridge, MA: MIT Press.

Dennett, D.C. (2007). What RoboMary knows. In Alter, T. and Walter, S. (eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, pp. 15–31. Oxford: Oxford University Press.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.

Farrell, B.A. (1950). Experience. *Mind* **59**, 170–198.

Frege, G. (1892). On sense and reference. In Geach, P. and Black, M. (eds) *Translations from the Philosophical Writings of Gottlob Frege*, pp. 56–78. Oxford: Blackwell, 1970.

Horgan, T. (1984). Jackson on physical information and qualia. *Philosophical Quarterly* **34**, 147–152. Reprinted in Ludlow, P., Nagasawa, Y., and Stoljar, D. (eds) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pp. 301–308. Cambridge, MA: MIT Press, 2004.

Horgan, T. and Tienson, J. (2001). Deconstructing new wave materialism. In Loewer, B. (ed.) *Physicalism and Its Discontents*, pp. 307–318. Cambridge: Cambridge University Press.

Jackson, F.C. (1982). Epiphenomenal qualia. *American Philosophical Quarterly* **32**, 127–36. Reprinted in Ludlow, P., Nagasawa, Y., and Stoljar, D. (eds) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pp. 9–50. Cambridge, MA: MIT Press, 2004.

Jackson, F.C. (1986). What Mary didn't know. *Journal of Philosophy* **83**, 291–295. Reprinted in Ludlow, P., Nagasawa, Y., and Stoljar, D. (eds) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pp. 51–56. Cambridge, MA: MIT Press, 2004.

Jackson, F.C. (1995). Postscript. In Moser, P.K. and Trout, J.D. (eds) *Contemporary Materialism*, pp. 184–189. London: Routledge. Reprinted in Ludlow, P., Nagasawa, Y., and Stoljar, D. (eds) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pp. 409–415. Cambridge, MA: MIT Press, 2004.

Jackson, F.C. (1998a). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

Jackson, F.C. (1998b). Postscript on qualia. In Jackson, F.C. *Mind, Method, and Conditionals*, pp. 76–79. London: Routledge. Reprinted in Ludlow, P., Nagasawa, Y., and Stoljar, D. (eds) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pp. 417–420. Cambridge, MA: MIT Press, 2004.

Jackson, F.C. (2003). Mind and illusion. In O'Hear, A. (ed.) *Minds and Persons* (Royal Institute of Philosophy Supplement 53), pp. 251–271. Cambridge: Cambridge University Press. Reprinted in Ludlow, P., Nagasawa, Y., and Stoljar, D. (eds) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pp. 421–442. Cambridge, MA: MIT Press, 2004.

Jackson, F.C. (2004). Foreword. In Ludlow, P., Nagasawa, Y., and Stoljar, D. (eds) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pp. xv–xix. Cambridge, MA: MIT Press.

Jackson, F.C. (2005a). Consciousness. In Jackson, F.C. and Smith, M. (eds) *The Oxford Handbook of Contemporary Philosophy*, pp. 310–333. Oxford: Oxford University Press.

Jackson, F.C. (2005b). The case for *a priori* physicalism. In Nimtz, C. and Beckermann, A. (eds) *Philosophy—Science—Scientific Philosophy: Main Lectures and Colloquia of GAP* 5, *Fifth International Congress of the Society for Analytical Philosophy, Bielefeld,* 22–26 *September* 2003, pp. 251–265. Paderborn: Mentis.

Jackson, F.C. (2006). On ensuring that physicalism is not a dual attribute theory in sheep's clothing. *Philosophical Studies* **131**, 227–49.

Jackson, F.C. (2007). The knowledge argument, diaphanousness, representationalism. Alter, T. and Walter, S. (eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, pp. 52–64. Oxford: Oxford University Press.

Kirk, R. (2006). Zombies. In Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2006 edition), http://plato.stanford.edu/archives/win2006/entries/zombies/.

Kripke, S.A. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

LeDoux, J.E. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon and Schuster.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* **64**, 354–361.

Levine, J. (1993). On leaving out what it's like. In Davies, M. and Humphreys, G.W. (eds) *Consciousness: Psychological and Philosophical Essays*, pp. 121–136. Oxford: Blackwell.

Levine, J. (2007). Phenomenal concepts and the materialist constraint. In Alter, T. and Walter, S. (eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, pp. 145–166. Oxford: Oxford University Press.

Lewis, D. (1988). What experience teaches. *Proceedings of the Russellian Society*. Sydney: University of Sydney. Reprinted in Lycan, W.G. (ed.) *Mind and Cognition: A Reader*, pp. 499–518. Oxford: Blackwell, 1990.

Loar, B. (1997). Phenomenal states (revised version). In Block, N. Flanagan, O., and Güzeldere, G. (eds). *The Nature of Consciousness: Philosophical Debates*, pp. 597–616. Cambridge, MA: MIT Press. Reprinted in Ludlow, P., Nagasawa, Y. and Stoljar, D. (eds) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pp. 219–239. Cambridge, MA: MIT Press, 2004.

Lodge, D. (2001). *Thinks …* London: Secker and Warburg.

Ludlow, P., Nagasawa, Y. and Stoljar, D. (eds) (2004). *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Cambridge, MA: MIT Press.

McGinn, C. (2004). *Consciousness and Its Objects*. Oxford: Oxford University Press.

McLaughlin, B.P. (2005). *A priori* versus *a posteriori* physicalism. In Nimtz, C. and Beckermann, A. (eds) *Philosophy—Science—Scientific Philosophy: Main Lectures and Colloquia of GAP* 5, *Fifth International Congress of the Society for Analytical Philosophy, Bielefeld,* 22–26 *September* 2003, pp. 267–285. Paderborn: Mentis.

Merikle,P.M., Smilek, D., and Eastwood, J.D. (2001). Perception without awareness: Perspectives from cognitive psychology. *Cognition* **79**, 115–134.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review* 83, 435–450. Reprinted in Nagel, T. *Mortal Questions*, pp. 165–180. Cambridge: Cambridge University Press, 1979. Also reprinted in Block, N. Flanagan, O., and Güzeldere, G. (eds) *The Nature of Consciousness: Philosophical Debates*, pp. 519–527. Cambridge, MA: MIT Press, 1997.

Nemirow, L. (1980). Review of *Mortal Questions* by Thomas Nagel. *Philosophical Review* **89**, 473–477.

Nemirow, L. (2007). So *this* is what it's like: A defense of the ability hypothesis. In Alter, T. and Walter, S. (eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, pp. 32–51. Oxford: Oxford University Press.

Nida-Rümelin, M. (1995). What Mary couldn't know: Belief about phenomenal states. In Metzinger, T. (ed.) *Conscious Experience*, pp. 219–441. Paderborn: Mentis. Reprinted in Ludlow, P., Nagasawa, Y. and Stoljar, D. (eds) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pp. 241–267. Cambridge, MA: MIT Press, 2004.

Nida-Rümelin, M. (2002). Qualia: The knowledge argument. In Zalta, E.N. (ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2002 edition), http://plato.stanford.edu/archives/fall2002/entries/qualia-knowledge/.

Papineau, D. (2002). *Thinking About Consciousness*. Oxford: Oxford University Press.

Papineau, D. (2007). Phenomenal and perceptual concepts. In Alter, T. and Walter, S. (eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, pp. 111–144. Oxford: Oxford University Press.

Quine, W.V.O. (1953). *From a Logical Point of View*. Cambridge, MA: Harvard University Press.

Rosenthal, D.M. (1991). The independence of consciousness and sensory quality. In Villanueva, E. (ed.) *Philosophical Issues, Volume* 1: *Consciousness*, pp.15–36. Atascadero, CA: Ridgeview. Reprinted in Rosenthal, D.M. *Consciousness and Mind*, pp. 135–148. Oxford: Oxford University Press, 2005.

Rosenthal, D.M. (2005). *Consciousness and Mind*. Oxford: Oxford University Press.

Russell, B. (1927). *The Analysis of Matter*. London: Kegan Paul.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Ryle, G. (1954). *Dilemmas: The Tarner Lectures* 1953. Cambridge: Cambridge University Press.

Seager, W. and Bourget, D. (2007). Representationalism about consciousness. In Velmans, M. and Schneider, S. (eds) *The Blackwell Companion to Consciousness*, pp. 261–276. Oxford: Blackwell.

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

Stoljar, D. (2001). Two conceptions of the physical. *Philosophy and Phenomenological Research* **62**, 253–281. Reprinted in Ludlow, P., Nagasawa, Y. and Stoljar, D. (eds) *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pp. 309–331. Cambridge, MA: MIT Press, 2004.

Stoljar, D. (2004). The argument from diaphanousness. In Ezcurdia, M., Stainton, R. and Viger, C. (eds) *New Essays in the Philosophy of Language and Mind (Supplementary Volume of the Canadian Journal of Philosophy)*, pp. 341–390. Calgary: University of Calgary Press.

Stoljar, D. (2005). Physicalism and phenomenal concepts. *Mind and Language* **20**, 469–494.

Stoljar, D. (2006). *Ignorance and Imagination: The Epistemic Origin of the Problem of Consciousness*. Oxford: Oxford University Press.

Stoljar, D. and Nagasawa, Y. (2004). Introduction. In Ludlow, P., Nagasawa, Y. and Stoljar, D. (eds). *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*, pp. 1–36. Cambridge, MA: MIT Press.

Sturgeon, S. (1994). The epistemic view of subjectivity. *Journal of Philosophy* **91**, 221–235.

Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press.

Tye, M. (1999). Phenomenal consciousness: The explanatory gap as a cognitive illusion. *Mind* **108**, 705–725. Reprinted in Tye, M. *Consciousness, Color, and Content*, pp. 21–42. Cambridge, MA: MIT Press, 2000.

Tye, M. (2000). *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.

Velmans, M. and Schneider, S. (eds) (2007). *The Blackwell Companion to Consciousness*. Oxford: Blackwell.

Watkins, M. (1989). The knowledge argument against 'the knowledge argument'. *Analysis* **49**, 158–160.

Weiskrantz, L. (1986). *Blindsight: A Case Study and Implications*. Oxford: Oxford University Press.

Weiskrantz, L. (1997). *Consciousness Lost and Found*. Oxford: Oxford University Press.