Australasian Journal of Philosophy

ISSN: 0004-8402 (Print) 1471-6828 (Online) Journal homepage: https://www.tandfonline.com/loi/rajp20

# Physicalism, consciousness and the antipathetic fallacy

David Papineau

**To cite this article:** David Papineau (1993) Physicalism, consciousness and the antipathetic fallacy, Australasian Journal of Philosophy, 71:2, 169-183, DOI: 10.1080/00048409312345182

**To link to this article:** https://doi.org/10.1080/00048409312345182

Published online: 02 Jun 2006.

*Australasian Journal of Philosophy*
Vol. 71, No. 2; June 1993

# PHYSICALISM, CONSCIOUSNESS AND THE ANTIPATHETIC FALLACY

David Papineau

## I. Introduction

In this paper I want to explain, from a physicalist point of view, why so many people are persuaded that consciousness is non-physical.

I take there to be good arguments, stemming from the need to integrate conscious events into the causal workings of the world, for identifying conscious states with physical states, and in what follows I shall take these arguments as read. At the same time there is no doubt that many people have strong intuitions that consciousness cannot possibly be physical. My aim will be to explain how these intuitions arise, and why they do not discredit physicalism.

I shall start with Frank Jackson's anti-physicalist 'knowledge argument' (Jackson [4], [5]) and explore the appropriate physicalist response. This will lead me to the identification of a fallacy, which I shall call the 'antipathetic fallacy', and which I shall argue seduces us into thinking of consciousness as something distinct from the physics of the brain. I shall conclude by showing how the identification of this fallacy enables us to stop asking for certain kinds of explanation of consciousness.

I shall follow common practice[1] in this area by using the term 'physicalism' broadly, to refer not only to the thesis that conscious states are strictly identical with states of the sort studied by physicists, but also to the thesis that they are identical with second-order or higher-order states, such as functional states, which are in turn realized by strictly physical states. The differences between these different kinds of broadly physical states will not matter here, since the arguments at issue between physicalists and their opponents arise in just the same way whichever kind of 'physical state' conscious states are identified with.[2]

## II. Jackson's Knowledge Argument

Consider Frank Jackson's story of Mary (Jackson [4],[5]). Mary is an expert on the psychology and physiology of human colour vision. Mary knows everything there is to know about what goes on in the brain when humans see red, say. However, Mary has always lived in a restricted black-and-white environment. She has never

[1] See Horgan [3, pp.147-148], Tye [14, p.1].

[2] What is more, they also arise in the same way between most non-physicalist 'objectivists' about mental states and their subjectivist opponents. Even so, I shall talk specifically about physicalism, since I think there are good arguments for preferring physicalism to other kinds of objectivism. Still, most of what follows should be of interest to objectivists in general and not just to physicalists.

herself seen anything red. Then one day she is presented with a red object. She then has the experience of seeing something red. And as a result she *learns something she didn't know before*. She now knows about the phenomenal nature of red colour experiences, when before she was ignorant of this. Remember, however, that Mary had always possessed complete physical information about colour experiences. So it seems to follow that there are items of information about experience — subjective facts, first-person facts, phenomenal facts — that must be omitted by any physicalist account.

The appropriate physicalist response to this argument is to admit that there are indeed before-and-after differences in Mary, consequent on her having had her first experience of red, but to deny that these involve her becoming acquainted with *some subjective feature* of colour experience. There are other ways of construing the changes in Mary, which do not require the postulation of such subjective facts, and which do not therefore imply that a physicalist account of experience must be incomplete.

In the next two sections I shall consider some of the physicalistically acceptable ways in which Mary changes as a result of seeing red. This will involve retreading some relatively familiar philosophical ground. But my aim is not just to block Jackson's argument — other philosophers, referred to below, have already shown how to do that — but rather to point to a striking common feature of the experientially-produced changes in Mary, namely, that they all yield ways of thinking about experiences that deploy versions of those same experiences. In section V I shall argue that this feature is responsible for the widespread intuition that consciousness cannot be physical.

In section III I shall consider how Mary acquires new *recreative* and *introspective* abilities. In section IV I shall consider her acquisition of new *concepts* of experience. But first let me deal with the most obvious way in which Mary changes as a result of her new experience. She has now *had* the experience of seeing red, whereas before she hadn't. Note that this change in itself raises no difficulty for physicalism. Physicalists think that conscious experiences are identical with certain physical events in the brain. So physicalists can simply say that the difference between the later Mary, who has experienced red, and the earlier Mary, who had not, is simply that certain physical events, namely those which constitute experiences of red, have now occurred in Mary.

Perhaps it is worth pausing briefly on this point, obvious as it is, since it is central to the physicalist view of conscious experience. Physicalism does not deny that there are conscious experiences, nor indeed that 'that it is like something to have them'. The claim is only that this is nothing different from what it is to *be* a physical system of the relevant kind. Of course there is something it is like to experience pain, or to see red, or to taste cheese. And such things are highly important, especially for the subjects of those experiences. But, insists the physicalist, they are not non-physical things. What makes it like that for you is that you *are* you, that is, that you are a physical system of a certain sort. If you were physically different in the relevant respects, things would be different for you.

## III. Recreation and Introspection

Let us now turn to some more interesting before-and-after differences in Mary. To begin with, there is the point that afterwards Mary is able to *recreate* the experience of seeing red, in imagination and memory, whereas before she couldn't. Mary could of course always imagine, in the third-person, so to speak, *that* somebody else was seeing red, in the sense that she could imagine such-and-such physiological or behavioural occurrences in that person. And, similarly, she was always able to remember, in the third-person again, *that* somebody had seen red. But now she has a new ability, the ability to imagine or recall *having* the experience itself, from the inside, as it were. She can now *relive* the experience, as opposed to just thinking about it.

An anti-physicalist account of this change will appeal to Mary's new knowledge of a non-physical fact. When Mary experiences red, on this account, she discovers that the experience has a characteristic phenomenal feature P. And then, because she has this new knowledge, she can imagine the experience by entertaining the thought that someone has an experience with property P. Similarly, she is now able to recall the experience by remembering that she herself had an experience with property P.

Physicalists will offer an alternative account. Suppose that the kind of imagination and memory at issue depends on the brain literally *recreating* a version of the experience being imagined or remembered. That is, suppose that first-person imagination or memory requires that brain be in a state which is similar to the state constituting the original experience. It won't be exactly the same state, since imagining or recalling a pain is different from having a pain. But it could well be a similar state, a kind of faint replica, which would fit with the fact that an imagined or remembered pain shares to some slight extent the unpleasantness of a real pain.

This alternative suggestion yields as good an explanation of the fact that you can only imagine or recall experiences you have previously undergone as the theory which postulates new knowledge of phenomenal property P. For it seems highly plausible that the brain's ability to recreate an experience depends, as a matter of empirical fact, on its having at some time had an original version of that experience, to give it, so to speak, the mould from which to make the replicas.[3]

Moreover, this alternative account of Mary's new ability is clearly quite consistent with a physicalist account of conscious experiences. For on this alternative account the difference produced in Mary by her original experience of seeing red is not that she acquires some new item of knowledge, but simply that she can now *do* something she could not do before, namely, recreate that experience in imagination

[3] An immediate qualification is needed here. For we can obviously imagine *complex* experiences, like seeing a unicorn, as long as we've previously experienced the elements separately. And we can perhaps imaginatively extrapolate to *intermediate* experiences, like imagining a colour which is spectrally between others we have previously experienced. But these possibilities are clearly consistent with the general thesis that the brain needs to acquire the materials for the replicas from previous experiences, and so in accord with the fact that we can't imagine experiences of a radically unfamiliar kind, like seeing colours at all, or echolocating, until we have actually had those experiences.

and memory.[4] The earlier account, which attributed new knowledge of phenomenal property P to Mary, implied that her previous third-person information about the experience left something out. However, since the new account does not credit Mary with any such new knowledge, there is now no implication that a physicalist account of conscious experience is incomplete.

Some readers may feel that this physicalist account of first-person imagination and memory is an *ad hoc* theory whose only attraction is that it saves physicalism. But this would be unjust. For the account also has the positive virtue, noted in passing above, of offering some explanation of why an imagined or remembered experience resembles the original experience itself — namely, that such imaginings and rememberings literally involve a copy of the original experience.[5]

D.H. Mellor uses the term 'secondary' to refer to this kind of copied experience, the kind of experience which occurs when we recreate in imagination or memory those primary experiences we have previously undergone.[6] The existence of such secondary experiences which resemble their primary versions will be central to my eventual explanation of the antipathetic fallacy.

Of course this talk of 'resemblance' between secondary and primary experiences needs further elaboration, both to specify what kind of replication is involved, and to explain how the resulting replicas mimic the original experiences in our cognitive workings. But I take it to be uncontentious that there is some phenomenon of resemblance here, and that the model of 'secondary' replicas of primary experiences offers a promising route to an explanation.

So far I have considered the new recreative powers of imagination and recall produced by Mary's first experience of seeing red. Another such before-and-after change is that Mary acquires a new *introspective* power to reidentify that experience when she has it again. Mary of course always had the ability to recognize 'from the outside' when somebody was seeing red, from environmental or behavioural or physiological evidence. But now she has a new ability, to recognize, by direct introspection, that she herself is seeing red.

Again, one possible explanation of this new first-person ability would be that Mary discovers that the experience of seeing red has phenomenal property P, and that as a result she can now pick out experiences with property P as instances of seeing red. But, as before, this is not the only possible explanation of the new ability. For we can suppose instead that Mary simply acquires a non-conceptual 'template',

[4] The view that Mary acquires new abilities rather than new knowledge is urged in Lewis [6] and Nemirow [12].

[5] Note that the alternative non-physicalist account, in terms of phenomenal property P, does nothing at all to explain why the exercise of our recreative abilities should in some sense make us re-experience the original mental state. Thinking of or remembering something as an event with some property P can in general have any experiential nature, or none at all. Of course, it could be argued that, in the particular case of some *phenomenal* property P, thinking of or remembering an event with that property involves recreating in your brain a copy of the experience characterized by the property. But, once this last move is made, then it becomes unnecessary to bring in the phenomenal property P to explain Mary's new imaginings and memories in the first place — for now we can simply explain these imaginings and memories directly, by appealing her recreative abilities.

[6] Mellor [10, p.11].

in David Lewis' phrase,[7] which can then be compared directly with further experiences, and cause Mary to believe that she is experiencing red again. She doesn't arrive at this belief by noting that the experience has property P, and concluding that it is an experience of seeing red. There is simply a mechanism in her brain which compares the experience with the template and yields this belief directly.

As with the 'secondary experience' account of imagination and memory, the 'template' account of introspective recognition both yields a plausible account of why we should need the original experience in order to acquire the recognitional ability (namely, because the brain needs the original to have the materials from which to form the template),[8] and remains consistent with physicalism (since it doesn't explain Mary's new recognitional ability by attributing knowledge of phenomenal property P to her, but simply by postulating a new mechanism in her brain).

It is an interesting further hypothesis that the same cognitive operations may be involved both in recreative and in recognitional abilities. Perhaps the brain uses the processes constituting our 'secondary experiences' themselves as the 'templates' by which it classifies new experiences: that is, perhaps its mechanism for recognizing such new experiences is simply to compare them with the replicas which are activated in imagination and recall. It does not seem inevitable that things should work like this: there is no contradiction in the idea of beings who could classify new experiences by some template process, and yet lacked the ability to recreate those experiences in imagination or memory; and perhaps it is even possible for there to be beings who could recreate experiences, but who lacked the second-order mental ability to classify them. But it seems clear that in human beings the two abilities always go together, and the natural explanation is that they do so because the same mechanism subserves both.

## IV. Concepts of Experience

The overall argument of the last section can be put as follows. In so far as Mary's first experience of red leads to her knowing something she didn't know before — leads to her 'knowing what the experience is like', if you want to put it that way — this new knowledge can all be construed as her knowing *how* to do something new, rather than as her knowing *that* anything new. There are indeed genuine changes produced by Mary's new experience. But these changes are all a matter of her acquiring new abilities — to recreate or recognize the experience — not of her forming any new kinds of judgements about the world.

But can this be the full story? Surely, many readers will feel, new experiences doesn't just give us the abilities described in the last section. They also enable us to think new thoughts. Once you've seen red, then can't you think of *that colour*, and

[7] Lewis [7, pp.131-132].

[8] Again a qualification is needed to accommodate the fact the we can recognize novel complex experiences, as long as their components have previously been experienced: in such cases the brain doesn't need an original complex experience to form a complex template, but only the originals of the component experiences to form templates of the components.

judge *it* to be vibrant, or threatening, or something everybody should experience at least once, in a way you couldn't before?

I agree. But I think that this too can be accommodated by physicalism. The important question for physicalism is whether new experiences lead to our knowing *about* any new features of the world. Physicalists need to deny this. But they can consistently allow that new experiences lead to our acquiring new *concepts* for thinking about those features. In Fregean terms, the change would be at the level of sense, not reference. Mary's *thinking* about the experience of seeing red would change, but what she was thinking *about* would be exactly the same thing as she used to think about when she was a scientist who had never herself seen red.[9]

In order to bring out this point, it will be helpful to switch examples slightly for a moment, and consider, not Mary, but Jane, let us call her. Jane has always shared Mary's black-and-white environment. But Jane is no expert on colour vision. Indeed she has never heard of such things as colours and of people experiencing them.

Then one day Jane sees something red. Unlike Mary, she does not have available any public concept of the visual events that take place in people when they are presented with red objects. Indeed she may not even realize that the sensation she is currently experiencing is caused by some observable feature of her environment. Yet we would surely expect Jane to be able thereafter to to form beliefs about that sensation, such as that it was vibrant, threatening, something everybody should experience it at least once, and so on.[10]

Mary, on the other hand, does have a public concept of the visual events that take place in people when they are presented with red objects. But, despite this, there seems no question but that Mary might acquire just the same new thoughts as Jane does after experiencing red for the first time. For imagine that, the first time Mary experiences red, she does not know what it is — she is simply not aware that the curious experience she is now having for the first time is the experience that is characteristically caused by red objects. In this case Mary will surely respond just like Jane, and start forming beliefs such as that this new experience is vibrant, threatening, and so on.

At first sight this might seem to substantiate Jackson's knowledge argument. Doesn't the fact that Mary follows Jane in forming new sorts of beliefs after her experience show that Mary's original set of physicalist beliefs must have left something out, namely, information about the subjective side of the experience? But this conclusion does not follow if, in line with my earlier suggestion, the novelty in Mary's beliefs lies at the level of sense rather than reference. And this of course is how the physicalist will diagnose the situation. Before Mary sees red, she has a 'third-person' concept of this experience. Afterwards she also has a 'first-person'

[9] Cf. Peacocke [13, pp.67-69].

[10] Isn't Jane ruled out by Wittgenstein's private language argument? Well, she'd better not be, if Wittgenstein's argument is any good, since Jane is clearly possible. I don't think there is in fact any tension here. I take the moral of Wittgenstein's argument to be that there must be room for error in people's judgements about their experiences, not that those judgements must necessarily be expressed in a language used by a community. And I see no reason to suppose that Jane cannot make mistakes about her own experiences.

concept. But they are concepts of the same thing. Mary is in the position of somebody who has thoughts about both Cicero and Tully, without realizing they are the same person.

A natural hypothesis about the structure of the new first-person concept acquired in common by Jane and Mary is that it involves a kind of exemplificatory reference by secondary experience. I earlier expressed the new belief formed by Jane and Mary as 'that experience is vibrant'. I now suggest that we take this construction at face value. Jane and Mary think: *THAT experience is vibrant*, accompanied by a secondary version of seeing red; they thereby secure reference to the experience of seeing red.[11] This account of first-person concepts of experience shows, as before, why you can't refer to experiences first-personally until you have had them, and it does so consistently with physicalism.

It is interesting to consider what will happen if Mary figures out that her new experience is characteristically occasioned by red objects. The natural upshot, assuming that Mary herself is a physicalist, it for her to conclude that she has two concepts with the same referent. And then, as with anybody who realizes this, the two concepts will tend to 'merge', each becoming merely an aspect of the unified concept with which she refers to the experience of seeing red.[12]

So, to sum up the argument of this section, once we have new experiences, we are led to form new sorts of beliefs about those experiences. But this does not show that we thereby come to refer to any distinctively subjectively phenomena. For the distinctive element in these beliefs need be nothing more that the deployment of first-person concepts, and, for all that has been said so far, there is no reason to suppose that such first-person concepts are not co-referential with third-person concepts of experience.

## V. The Antipathetic Fallacy

I expect that, despite everything I have said so far, many readers will feel strongly that it is a mistake to conclude that 'first-person' and 'third-person' concepts of experience refer to the same things. For my arguments in the last two sections will have done nothing to shake the widespread intuition that conscious experiences and brain states are as different as anything can be.

Let me summarize the state of play. So far I have considered the strength of the 'knowledge argument' *against* the physicalist identification of conscious experiences with brain states. I take the points made in the last two sections to show that this argument is ineffective.

Now add in the consideration, which I haven't argued for here, but which I mentioned at the beginning, that the need to integrate conscious states into the causal workings of the world gives us a strong motive for identifying them with physical states. I think that this, together with the ineffectiveness of the 'knowledge argu-

[11] This suggestion is central to the response made to Jackson's argument in Horgan [3].

[12] Of course, if Mary isn't a physicalist, then she will be disinclined to make this identification, and will no doubt maintain that the first-person concept she shares with Jane refers to a phenomenal attribute, whereas her scientific concept refers to a physical phenomenon.

ment', now gives us good reason to accept the physicalist view that conscious experiences are not distinct from brain states, and therefore good reason to reject any intuitions to the contrary. However, it would be foolish to deny that such intuitions exist. Such non-physicalist intuitions exert a strong pull on all of us, even on us physicalist philosophers who are committed to rejecting them. So in this section I want to offer the following diagnosis of these intuitions, with the intention of explaining why they arise even though they are mistaken.

In the previous two sections I have discussed a variety of ways in which we can focus mentally on conscious experiences, a variety of mental acts which refer to types of experience. These acts can be divided into two main categories: those 'third-person' acts which are possible prior to your having had the experience in question, and those 'first-person' acts which are only possible after you have had the experience. In the former category are all the mental acts Mary could perform before she saw red: her 'third-person' imaginings and memories of other people experiencing red; her non-introspective identification on behavioural or physiological grounds of certain events as experiences of seeing red; her 'third-person' beliefs, conjectures, and other propositional attitudes about the experience of seeing red. In the latter category are the 'subjective' analogues of all these mental acts: the 'first-person' imaginings and rememberings that involve internal recreation of an original experience; the introspective identifications of new experiences by direct comparison with a 'template'; the beliefs, conjectures, and other attitudes that can be formed by people like Jane whose concept of seeing red involves an element of ostension by internal exemplification.

The common feature of these latter 'first-person' acts, and what distinguishes them from the corresponding 'third-person' acts, is that they all deploy a secondary version of the experience being referred to. This is the reason, I have suggested, why the first-person acts are only possible after you have had the experience in question yourself. For it is only after you have had the experience that your brain will have the materials necessary to form secondary versions of that experience.

I think that this broad division between first-person ways of thinking about experience, which employ secondary versions which resemble those experiences, and third-person ways, which do not, is the source of the strong intuition that conscious experiences involve something more than the physics of the brain. For it is all too easy to conclude, when we reflect on the difference between these two categories of thought, that only the first-person thoughts really refer to experiences, while the third-person thoughts refer to nothing except physical states.

The route to this conclusion begins with the perfectly accurate observation that first-person thoughts include an experiential element which is absent from the third-person cases. First-person thoughts portray the relevant experience directly, so to speak, by giving the thinker a simulacrum, by recreating in the thinker a version of the experience being thought about. Third-person thoughts, on the other hand, do not do this, since they do not involve secondary experiences.[13]

[13] Or, if they do involve secondary experiences, as when we think about somebody being in pain, say, by thinking about the visual aspect of their behaviour or brain state, then they will be different secondary experiences, secondary versions of visual experiences, not secondary versions of pain experiences.

So there is a sense in which third-person thoughts indeed 'leave something out': they do not *give* us (versions of) the experience being referred to. And this observation can then easily lead to the further conclusion that third-person thoughts are about something different from first-person thoughts: where first-person thoughts refer to the experience itself, in all its conscious immediacy, third-person thoughts merely refer to the external trappings of the conscious event, the physical goings-on which accompany it.

But of course this last step is a fallacy. The fact that we do not *have* certain experiences when we think third-person thoughts does not mean that we are not *referring* to them. To make this move is to succumb to a species of the use-mention confusion: we slide from (a) third-person thoughts, unlike first-person thoughts, do not *use* (secondary versions) of conscious experiences to portray conscious experiences to (b) third-person thoughts, unlike first-person thoughts, do not *mention* conscious experiences. There is no reason, however, why third-person thought about experiences, like nearly all other thoughts about anything, should not succeed in referring to items they do not use.

I propose to call the above fallacy the 'antipathetic fallacy'. Ruskin coined the phrase 'pathetic fallacy' for the poetic figure of speech which attributes human feelings to nature ('the deep and gloomy wood', 'the shady sadness of a vale'). I am currently discussing a converse fallacy, where we refuse to recognize that conscious feelings inhere in certain parts of nature, namely, the brains of conscious beings.

Let me be specific about the target of this charge of fallacy. My target is not the explicit argument against physicalist views of consciousness offered by Jackson. I take the points made in the last two sections already to have shown what is wrong with Jackson's argument. Rather my target is a covert line of thought, whose fallaciousness is obvious once it is spelt out, but which I think has nevertheless seduced a great many thinkers into dualism: namely, the argument which moves from the true premise that third-person ways of thinking about conscious experiences do not use versions of those conscious experiences, to the false conclusion that those ways of thinking do not mention those conscious experience, but only physical states.

Let me also be specific about what I take the identification of this fallacy to explain. It is supposed to explain why many people *believe* that some mental states are distinctively *non-physical.* It is *not* supposed to explain why it is true that some physical states are distinctively *conscious.* This latter kind of question will be addressed in the final two sections of this paper, and I shall there agree that our ability to think about certain states in first-person ways does nothing to account for their possessing the distinctive inner light of consciousness — though I shall also argue there that the desire to account for such inner lights rests on a confusion. My present concern, however, is not to explain why the states we can think about in first person ways are distinctively conscious, but rather to explain why these states are widely taken to be non-physical.

Both Thomas Nagel, in a well-known footnote in 'What is it Like to be a Bat?' [11, pp.446], and William Lycan, in his book *Consciousness* [8, pp.76-77], briefly allude to versions of the fallacy I am concerned with. My treatment here enlarges on their remarks in two respects. First, both Nagel and Lycan focus specifically on

the contrast between first-person imagination of conscious experiences and the third-person perceptual imagination of the associated brain states: the contrast, for example, between imagining *having* a pain and imagining the *visual appearance* of the relevant parts of the sufferer's brain. This is certainly one example of the kind of contrast I am interested in, but this exclusive focus underemphasizes the extent of this contrast. For, as I have observed, the contrast between first-person and third-person modes of thought is not restricted to imagination, but also includes memory, introspective identification, and believing, desiring and other propositional attitudinizing. And even within the category of imagination, perceptual imagination is not the only kind of third-person imagination: if we can form non-perceptual beliefs and other propositional attitudes about brain states, as we surely can, then presumably we can imagine them non-perceptually too. (Nagel does mention 'symbolic imagination', but only to exclude it from his analysis.)

Second, neither Nagel nor Lycan emphasize the way that first-person modes of thinking about experiences deploy *secondary versions* of those experiences. Nagel does, it is true, say that first-person imaginings 'resemble' the experiences being imagined. But when he goes on to explain how the fallacy arises, his explanation, like Lycan's, is simply that first-person and third-person imaginings are independent mental acts, each of which can happen without the other, and that therefore we are inclined to conclude that they are about different things.[14] But this diagnosis fails to distinguish the antipathetic fallacy from all the other cases where different modes of thought about the same entity can create the impression that two different entities are being thought about. What is distinctive about the antipathetic fallacy, and what makes it so very seductive, is the fact that one set of ways of thinking about experiences — the first-person ways — involve versions of the experience itself, and so create the impression that the other ways of thinking about experiences — the third-person ways — *leave something out*. In general, when two different mode of thought create the impression that two things are being thought about (for example, Cicero and Tully), the illusion is easily enough dispelled on receipt of evidence that there is in fact only one referent. But in the mind-body case the impression of difference continues even in the face of any amount of such evidence, precisely because of the extra feature — the first-person use of secondary versions — that makes it seem as if the third-person modes of thought omit mention of the experience altogether.

## VI. Theories of Consciousness

So far I have argued that there are no effective arguments against the physicalist identification of conscious states with physical states, and that the admittedly strong intuitions which run counter to this view can be explained away. It may still seem to some readers, however, that there is a further obligation facing defenders of a physicalist view of consciousness. Do they not need to answer the challenge raised briefly in the middle of the last section, namely, to explain why some states are conscious and others not?

[14] Accordingly Lycan calls his version of the fallacy the 'stereoscopic fallacy'.

The obligation I am thinking of here is not just to provide physicalistically acceptable accounts of such specific conscious states as being in pain, seeing red, having an itch in your left finger, and so on. Let us suppose that physicalists can somehow specify which physical occurrences constitute each of these mental states. The current challenge is to explain what ties these different states together. Why are pains, itches, and so on, all conscious, while the states of stones, hydrogen atoms, and golf balls, are not?

Some philosophers of physicalist inclinations have proposed 'theories of consciousness' in answer to this kind of question. I have in mind the kind of theory which aims to identify a physicalistically acceptable characteristic common to all and only conscious states. Some such theories are based on assumptions drawn from everyday thought (for example, Armstrong [1, pp.92-99]); others appeal to the resources of cognitive science (for example, Dennett [2, ch.9]); and none, I think, commands universal assent.

However, we can leave the details of such theories to one side. For I am sure that many readers will feel that any theory of this kind will inevitably fail to address the philosophical question at issue. At best such a theory will simply specify some structural or other physically acceptable characteristic (A, say) which is coextensive with the class of states we are pretheoretically inclined to count as conscious.[15] But then we still seem to face the question: *why* does consciousness emerge in just those cases? And to this question physicalist 'theories of consciousness' seem to provide no answer.

I suspect that many philosophers regard the inability to answer this question as the fatal flaw in the physicalist approach to consciousness. Surely, they feel, any satisfactory philosophical view of consciousness ought to tell us why consciousness emerges in some physical systems but not others.

I think that physicalists should simply reject this question. For the question presupposes that there are two different features at issue, the physically acceptable characteristic A, and being conscious. We are then asked for an account of the relation between these properties, and in particular of why they are always found together. But the physicalist should simply deny that there are two properties here. Being conscious isn't something over and above having A, it just *is* having A.

The idea that being conscious just *is* having some physical state might seem intuitively implausible: surely the difference between conscious and non-conscious systems is something more than the difference between having and lacking some physical feature. But the defender of a physicalist theory of consciousness, while not denying that these intuitions exist, can account for them as a further manifestation of the antipathetic fallacy. The earlier sections of this paper were concerned with the thesis that specific conscious states, like seeing red, are identical with specific physical states; and I argued there that our strong contrary intuitions can be explained

[15] It is possible that there isn't any physicalistically acceptable necessary and sufficient criterion for consciousness: perhaps there is simply some kind of family resemblance between the behavioural and other functional features we take to evidence consciousness. In that case, some of the claims which follow should be phrased more circumspectly; but my substantial points would remain the same.

away as due to the antipathetic fallacy. I would now like to suggest that a generalized version of this fallacy is responsible for the intuition that any physicalist theory of consciousness will necessarily be incomplete

We can think of the general property of being conscious as standing to experiences like seeing red as determinable to determinate. Seeing red, being jealous, feeling cold, and so on, are the determinate states which have in common the determinable state of being conscious. And so, just as the antipathetic fallacy makes us think that such determinate states as seeing red are distinct from any specific physical states, so it makes us think that the determinable state of being conscious is similarly distinct from any more general physical state. We are inclined to think of the determinable feature as a kind of generalized non-physical light, which stands to the non-physical features of particular experiences, as, say, the property of being illuminated as such stands to being illuminated with red light. But we shouldn't. Just as it is a mistake to think of experiencing red as something additional to the relevant physical property, so it is a mistake to think of being conscious as an extra inner light, over and above the physical feature A.

Once we fully free ourselves from the seductive 'inner light' picture of consciousness, and take seriously the idea that being conscious may literally be identical with some physical A, then we should stop hankering for any further explanation of *why* physical state A yields consciousness. Consider this parable. Suppose that there are two groups of historians, one of which studies the famous American writer Mark Twain, while the other studies his less well-known contemporary. Samuel Clemens. The two groups have heard of each other, but their paths have tended not to cross. Then one year they both hold symposia at the American Historical Association, and late one night in the bar of the Chicago Sheraton the penny drops, and they realize they have both been studying the same person. At this stage there are plenty of questions they might ask. Why did this person go under two names? Why did it take us so long to realize Mark Twain and Samuel Clemens were the same person? But it doesn't make sense for them to ask: why *were* Mark Twain and Samuel Clemens the same person? If they were, they were, and there's an end on it.[16]

Similarly, the defenders of a physicalist theory of consciousness can say, with consciousness and the physical property A. Such physicalists will take themselves to have discovered that consciousness and A are the same property. So they will allow that we can sensibly ask why there should be different concepts of this property, and why it took us so long to realize that they stand for the same thing; and indeed they can answer these questions, by explaining that there are ways of referring to conscious phenomena that use secondary versions of those phenomena, and ways that don't, and that this in itself makes it easy to succumb to the antipathetic fallacy of supposing that different things are being referred to. But, the physicalists will continue, there is no further question of *why* consciousness is always present when physical property A is. If they really are the same thing, then we can't explain *why* they are the same thing. Somebody who feels there is still a question here has

[16] Ned Block offered this story to me; I don't know where it originated.

simply failed properly to grasp the thesis that consciousness is identical with a physical property.

## VII. Life and Consciousness

It may seem to some readers as a physicalist theory of consciousness will come close to denying the existence of consciousness. But that would be a mistake. It doesn't deny consciousness, just a certain conception of consciousness.

It denies that consciousness is some kind of extra inner light, some further non-physical property which exists over and above any physicalistically specifiable property. But this is quite consistent with holding that consciousness is a real property which distinguishes some kinds of systems from others. This combination of views requires only that we accept that consciousness is identical with some property which is specifiable in a physicalistically acceptable way.

An analogy may be helpful here. In the nineteenth century there was a heated theoretical debate about the essence of *life*. The participants had a satisfactory enough working notion of life: they agreed about which kinds of behaviour and physical organization are characteristic of life, and in consequence were clear about where the line should in practice be drawn. Everything from humans to microbes are alive, while planets and pebbles are dead. (Perhaps questions could be raised about large organic molecules, or crystals; but the penumbra of vagueness was not wide.)

Still, despite this wide degree of agreement on the nature of life, nineteenth-century thinkers took there to be a further question. *Why* are these systems alive? What mysterious power animates them? And why is this power present in certain cases, such as trees and oysters, and not in others, like volcanos and clouds?

These questions have disappeared from active debate. Nobody nowadays asks why living systems are alive. Everybody is happy to agree that the difference between living and non-living systems is simply having a certain kind of physical organization (roughly, we would now say, the kind of physical organization which fosters survival and reproduction).

The explanation of the nineteenth-century debate, and of its subsequent disappearance, was that it was premised on the notion that living systems were animated by the presence of a special substance, a vital spirit, or *elan vital*, which was postulated to account for those features of living systems, such as generation and development, which were though to be beyond physical explanation. Of course, if you do believe in such a vital spirit, then you will want to know about its nature, and why it arises in certain circumstances and not others.

However, nobody nowadays believes in vital spirits any more, not least because it is now generally accepted that the characteristic features of living systems can in principle all be accounted for in physical terms. In consequence, it no longer makes sense to puzzle about *why* living systems are alive. To be alive is just to be a physical system of a certain general kind. There isn't any extra property present in living systems, over and above their physical features, which distinguishes them from non-living systems. So we have stopped asking questions which presuppose such an

extra property.

I recommend that we do the same with consciousness. The apparently nagging question, 'Why does consciousness arise in certain physical systems?', is premised, I claim, on the assumption that consciousness is some extra feature, over and above physical characteristics. But if we accept, as I have argued, that there is no reason to view consciousness in this way, then we ought therewith to stop asking why consciousness is present in the relevant kind of physical system.

Of course the parallel is not complete. In the case of life, the motivation for postulating an *elan vital* is purely explanatory, a desire to find a cause for phenomena which do not appear to be physically explainable. In the case of consciousness, by contrast, there is also the extra pressure of the antipathetic fallacy. Still, this doesn't affect the point. There may be extra causes for our thinking of consciousness as non-physical, which don't apply to life. But once we recognize that it is physical, we should do what we did with life, namely, stop asking why it arises in the right physical circumstances.

One last point about the analogy with life. Note that the rejection of an *elan vital* does not mean that there is no life. There may be nothing special about living systems except a certain kind of physical organization. But this does not mean that the difference between being alive and not being alive is not real. The postulation of an *elan vital* was simply one theory about the nature of life. We can reject this theory, and yet still uphold, as we do, the distinction between living and inanimate systems.

A similar point applies to consciousness. We should reject the theory that consciousness involves an inner light extra to facts of physical organization. But we can reject this theory without rejecting consciousness. Even if consciousness is just a kind of abstract physical organization, the difference between being conscious and not being conscious can still be perfectly real.[17]

*King's College ,London*

Received March 1992
Revised October 1992

[17] I would like to thank Tim Crane, Andrew Jack, Keith Hossack, Lucy O'Brien, Gabriel Segal, Barry Smith, Scott Sturgeon and the readers and editorial panel of the *Australasian Journal of Philosophy* for helpful comments on this paper.

REFERENCES

1. D.M. Armstrong, *A Materialist Theory of the Mind* (London: Routledge and Kegan Paul, 1968).
2. D. Dennett, *Brainstorms* (Hassocks, Sussex: Harvester, 1979).
3. T. Horgan, 'Jackson on Physical Information and *Qualia*', *Philosophical Quarterly* 32 (1984) pp.147-152.
4. F. Jackson, 'Epiphenomenal Qualia', *Philosophical Quarterly* 32 (1982) pp.127-136.
5. F. Jackson, 'What Mary Didn't Know', *The Journal of Philosophy* 83 (1986) pp.291-295.
6. D. Lewis, 'What Experience Teaches' *Procedings of the Russellian Society of Sydney University* (1988); reprinted in [9, pp.499-519].
7. D. Lewis, 'Postscript to "Mad Pain and Martian Pain"', in his *Philosophical Papers, Volume 1* (Oxford: Blackwell, 1983) pp.130-132.
8. W. Lycan, *Consciousness* (Cambridge, MA: MIT Press, 1987).
9. W. Lycan (ed.) *Mind and Cognition: A Reader* (Oxford: Blackwell, 1990).
10. D.H. Mellor, 'Nothing Like Experience', *Proceedings of Aristotelian Society,* 93 (1993) pp.1-16.
11. T. Nagel, 'What is it Like to be a Bat?', *Philosophical Review* 83 (1974) pp.435-450.
12. L. Nemirow, 'Physicalism and the Cognitive Role of Acquaintance', in [9, pp.490-499].
13. C. Peacocke, 'No Resting Place: A Critical Notice of *The View from Nowhere*', *The Philosophical Review* 98 (1989) pp.65-82.
14. M. Tye, 'The Subjective Qualities of Experience', *Mind* 95 (1986) pp.1-17.