Are mental states reducible to brain states?

or

The quale is dead: Long live the quale!

CONTENTS

Are mental states reducible to brain states?
or
The quale is dead: Long live the quale!

Abstract.

Each of the various philosophical positions on the mind-body problem has grown out of the perceived shortcomings of one or more of its predecessors. One fertile source of aggravation to many of the *-isms* has been the problem of qualia: the ostensibly irreducible, qualitative character of many of our mental states. An argument is presented here that solves the qualia problem within the context of an otherwise functionalist theory of mind. The proposed solution is unusual in that it both resolves the mystery of qualia and allows it to stand: there is in this an implied reconciliation between functionalism and epiphenomenalism. Along the way, two other issues are discussed: the putative distinction between functionalism and the identity thesis, and the role of such terms as 'reduction', 'explanation' and 'meaning' in the science of psychology.

1) Introduction

Questions about whether mental states can be reduced to brain states are significantly dependant on our concept of "reduction" in science, and reduction, in turn, has to do with what it means for us to "understand". There is obviously a potential for some tangling here: neither understanding nor reduction are absolute concepts; they are part of the functioning of the human cognitive system, and if it turns out that there are certain aspects of the world that seem impossible for us to understand or give a reductive explanation for, it may be that the fault lies not with some peculiarity of the world, but rather with our capacity to understand the world. This doesn't amount to the (trivial) claim that we need to develop the sophistication of our concepts in order to be able to understand some particularly knotty scientific and philosophical problems. Much more important, it may be that there are aspects of the world that can *never* be explained.

These ideas are expanded later to form a proposal for dealing with the question whether mental states are reducible to brain states. In the meantime, sections 2, 3 and 4, below, form a brief catalogue of the attempts that have been made to resolve this problem in the past.

2. Dualism

The essence of dualism is the assertion that mental states and brain states are distinctly different types of thing. There are two main versions of this claim; substance-dualism and property-dualism.

2.1) Substance Dualism

For a substance dualist mental states are distinguished by being states of a non-physical substance that is entirely different from normal matter. The mind is fabricated from a stuff that physics can have no truck with. In the original formulation of this idea by Descartes the mental stuff was also non-physical in that it was not constrained to be located at any definite point in space or time, and this was what gave it its unique character, making it recognisably different from normal matter.

Substance dualism suffers from many problems, but the most important is that it is difficult to square the idea of a mind that has no spatial or temporal location with the observation that distinct minds seem to be attached to particular bodies which certainly do not have the freedom to be everywhere at all times. Since the mind is entirely divorced from the physical world it is also a wonder that a) it can send signals down to control the body, b) it can get information about what is happening to the body and c) it shows a systematic array of deficits in response to damage in different parts of the brain.

2.2) Property Dualism

The tenets of property dualism hold a possible remedy for these difficulties, because here the mind is identified not as a non-physical *thing,* but rather as an extraordinary *property* of the brain. Brains give rise to minds, so the story goes, because of their complexity and organisation (and perhaps on account of some other factors, such as their neural constitution), so there is no need to worry about why minds are always seen hanging around in the vicinity of brains. It is worth emphasising that what makes this a dualist position is that the mind is supposed not to be predictable from any known physical properties of the brain. Property dualism itself divides into two (major) schools, according to the causal role assigned to the mind:

2.2 a) Epiphenomenalism

For an epiphenomenalist the mind cannot affect the brain in any way. It is just a "side-effect" of the functioning of the brain. One apparently strange consequence of this way of looking at things

is that volition, which is surely in the domain of the mental, is stripped of its purpose. Volition is the mental state that is the source of our actions, but if mental states do not have any effect on things in the physical domain, how could our wanting to do something with a part of our body ever actually cause it to happen?

2.2 b) Interactionism

Interactionism allows that the mind could be in a state of two-way communication with the brain, so the problems that plague epiphenomenalism don't arise.

3) Monism

Apart from all the specific difficulies that there are with the various forms of dualism, there is also a more general criticism to the effect that if mental states can be satisfactorily understood as arising out of some physical properties of the brain, then why bother to postulate something non-physical to explain them? Monism embraces a number of schools of thought that try to explain the mind in this way.

3.1) Physicalism

According to physicalism, mental states and brain states are one and the same. There are various forms of this thesis, according to whether the states in question are types or tokens: type-type physicalism asserts that for each type of mental event there is a corresponding type of event in the brain; type-token physicalism says that for each type of mental event there is a particular token of a brain event (one that happens at a particular place and time), and so on.

3.2) Eliminative Materialism

The eliminative materialist takes a simpler line on the problem of accounting for mental states. She would deny that that there is any sense in the notion of a mental state. These concepts will eventually be eliminated from our scientific understanding of the world, in just the same way that, for example, the notions of planetary epicycles, demonic possession and witches have been discarded by the physical sciences of the 20th century. Instead, there will be a perfectly satisfactory collection of terms to describe behavioural states, which refer only to events and states in the brain.

3.3) Functionalism

Physicalism is plagued by at least one important problem: if mental states and brain states are one and the same, then presumably anything made out of some form of matter that didn't include neural tissue would not be capable of having mental states. It could never be said of a computer, then, or of an extraterrestrial whose bodily processes were molybdenum-based, that they suffered pain, even in case they both had a psychology very similar to our own, and even though each might protest that she felt pain when you pulled out her diodes (or stuck a pin in her left fore-blob). This seems uncomfortably restrictive.

Functionalism offers the most congenial resolution of this difficulty. What characterises a mental state, according to the functionalist, are the causal connections that it has with other mental states and with the perceptual input and motor output states of the system. So, as long as the extraterrestrial or the computer had apparatus that was functionally equivalent to our own information processing system, it would be quite valid to allow that it had mental states in the way that we do, no matter what its physical constitution was.

At the present time, functionalism (or some variant thereof) seems to be in the ascendancy in philosophy and cognitive science, but that is not to say that it is without its critics. The next

section is devoted to a closer examination of functionalism's weak points.

4) Problems with functionalism

There are two outstanding problems which, it is claimed, functionalism has difficulty resolving. One concerns the subjective aspects of our experience and the other, the question whether a machine that was a functional analogue of a human mind could ever be said to be an "intentional object".

4.1) Qualia

There seems, to most people, to be something about the inner aspects of our experience that is not even addressed by a functionalist characterisation of mind. The term "qualia" is used to denote those qualitative aspects of some mental states that seem indescribably subjective and incommunicable; for example, the essence of a colour sensation or the quality of a pain sensation that makes it, so to speak, "horrible". Another perspective on this idea was given by Nagel (1974), when he asked "What is it like to be a bat?" - he points out that an objective description of the structure of another being's cognitive system, no matter how detailed, is powerless to convey exactly what it would feel like to be that being.

The orthodox functionalist position on this matter (Harman, 1982) is that both the concept of red and the quale associated with sensations of redness are defined by their functional relationships to other states. Whatever is required by way of relationships to other states and to input and output states, to make a given state a state of perceiving something red, will also be sufficient to ensure that a redness quale is associated with that state. This statement can be illustrated by two classic conundra known as the the "Qualia Inversion" and "Absent Qualia" problems.

4.2) Inverted qualia

Suppose that someone were as proficient at colour discriminations as you were, and made all the same reports as you when naming the colours of objects, but that this person actually experienced different colour sensations. So, when looking at a red object you would both say "red", but she would be experiencing the qualitative sensation that, if you were privy to it, would look like green. Correspondingly, you would compare notes correctly on a "green" object, but she would experience a sensation of red rather than green. Clearly, this "inversion" of your inner perceptions of colour could give rise to no outwardly observable differences in your behaviour; so, functionally, you would be described as both being in the same state when looking at a green object. This is, at first sight, a disturbing conclusion, because the functional description of your mental states would evidently not have captured everything that could be said about them.

What the orthodox functionalist position amounts to, in this case, is the bare assertion that qualia inversion simply cannot happen - if you describe things the same way as someone else, you both see the same qualia. This attempt to meet the objection leaves something to be desired.

4.3) Absent qualia

A related case is the "Absent Qualia" problem, wherein it is suggested that two functionally equivalent minds could nevertheless differ in that one could be entirely devoid of qualia. The orthodox functionalist would again say that this is possible in logical fact, but that in practice it does not actually happen. As Shoemaker (1975; 1981) points out, though, the argument is a little stronger in this situation because the assumed functional identity of the two beings, from the point of view of external observers, must surely imply that they would both report that they had qualia. In spite of White's (1986) reservations about Shoemaker's point, it has to be said that the notion of a qualia-deprived being that thought it did have qualia is a little hard to swallow. (Nevertheless, an argument very similar to this one will reappear below in the discussion of intentionality).

The absent qualia and inverted qualia problems can be used to classify some of the more refined versions of functionalism that depart from the "orthodox" doctrine given above. The terminology given here is due to White (1986).

4.4) Physicalist-functionalism

A physicalist-functionalist believes that the nature of one's qualia is determined by the physical substrate on which her (functionally characterised) cognitive system is realised. There is a distinction between type 1 and type 2 physicalist-functionalists (hereinafter "p-f"s) on the basis of the combination of absent and/or inverted qualia that are deemed coherently possible: the p-f(1) (Block, 1980) accepts that a suitable choice of substrate material could render a subject either quale-less or having inverted qualia (with respect to conspecifics), whilst the p-f(2) (Shoemaker, 1982) accepts that qualia inversion is possible, as a function of substrate, but that absent qualia are not, for the reasons earlier given. It has to be said that these attempts at a compromise between functionalism and physicalism have not had an unreserved welcome (White 1986).

4.5) Intentionality

In his paper "Minds, Brains and Programs", Searle (1980) set out an argument designed to show that whilst a machine could have all the outward signs of human-like mentality (i.e., allowing that it could be functionally equivalent to a human being), it could never be an intentional object. An intentional object is defined as one that could properly be the subject of a verb of propositional attitude. Propositional attitude verbs are ones such as "believe", "want", "wish", "desire" and so on, so what is being claimed here is that "robot" could never be a bona fide value for X in sentences of the form "X desires that p", "X knows that q", and so on. The substance of Searle's claim is that the machine can manipulate strings of symbols as symbols (that is, the machine can apply rules of syntax), but it has no access to the meaning of those symbols. It is in consequence of this absence of semantics that ascriptions of belief, etc. to machines makes no sense.

Searle builds his case on what amounts to an appeal to intuition in the context of a thought experiment. Briefly, the thought experiment in question is that a computer program designed to be intelligent (= a functional characterisation of a cognitive system) could be implemented on a (strange) machine consisting of an already intelligent system - i.e., a person (Block (1978) discusses a similar experiment involving a whole host of people (to wit, everyone in China) implementing a functional description of an intelligent system). Under these circumstances, Searle claims, the system of person + program might appear to be intelligent, but the person need have no knowledge of what the program she is following is actually doing. If the person, in this case, corresponds to the computer in the more normal case, then says Searle, we must conclude that the computer manipulates symbols but doesn't understand the meaning of what it says (or does). His final conclusion is that "intentionality" - the ability to understand meaning - is the special privilege of systems constructed from neural tissue.

This example of Searle's suffers from one obvious weakness of the sort that was pointed out above, in connection with qualia: the machine would, we presume, be able to have long and sophisticated discussions with us about "meaning", and what it involved, and it would *undoubtedly* (recall that Searle concedes that the machine could be in every respect functionally equivalent to ourselves) report that it understood the meaning of the terms and concepts that it used. It is alarming, to say the least, that any attempt would be made under those circumstances to explicitly deny that the machine could understand meaning.

But perhaps we don't need to credit Searle's example with even this much coherence: it seems quite clear to all but those unwilling to think the matter through that what Searle should conclude from his example is that the physical substrate of an intelligent system cannot be said to understand something even when the system as a whole evidently can. I might say of myself that I

know the meaning of so-and-so, but I would not for a moment try to claim that one of my neurons "knows" the same thing. It is quite reasonable to insist that a property like "knows p" can apply validly to a complex system whilst at the same time not being validly applicable to a component of the system (compare: "that piece of cumulo-nimbus is going to start raining soon" and "that raindrop in the piece of cumulo-nimbus there is going to start raining soon"....the second sentence may have poetic or metaphorical meaning, but not literal meaning). If Searle wishes to assert only that intentional terms cannot be used of intelligent-system-components, then no (orthodox) functionalist would wish to disagree. What is utterly without foundation is his conclusion that a machine (i.e. non-neural) system as a whole could not be said to have any grasp of meaning.

Ultimately, Searle relies on an appeal to intuition to carry his case. It is the *unreasonableness* of ascribing mentality to a system comprised of all of the people in China passing slips of paper to one another, or of one person alone in a room going through the tedious ritual of running a program by hand, that Searle would have us applaud. Alas, all one can say to this is that one person's argument from a priori plausibility is another person's specious assumption.

5) Qualia as ficticious concepts

The problem addressed by this essay is essentially an ontological one: what kind of thing is a mental state?; What status should qualia have in our scientific understanding of the world? One kind of answer that the question about qualia could receive has not been given much attention, and that is that they may be ficticious constructs - illusions caused by the mechanisms that we have available to us (sic.) for understanding the world.

Consider: when we look out at the world and try to understand it, we engage in a process of constructing concepts to model aspects of that world. Notice that one of the most important ways to build new concepts is to make combinations of old ones. And part and parcel of this conceptual development process is the game that involves picking concepts apart to look at the constituents or precursors: asking questions about what a thing "is" amounts to asking for an analysis of the concept of the thing at a more "basic" level of description (that is, using "simpler" or more "general" concepts). This can be the everyday, automatic answering of questions or it can be the systematic and rarified process of scientific reduction and explanation - arguably, these processes have a common foundation in the "picking apart" of concepts.

The mind obviously needs a facility for forming concepts to represent any and every detectable regularity in the perceived world. Since we need to have a good deal of information about ourselves, it is no surprise that the mind forms concepts to represent aspects of its own mechanism. But here there are concepts that are unlike others in some of their properties. What is it to know the difference between blue and yellow, for example? At bottom, when all contingent associations to the colours are stripped away, the difference is that if a certain input channel gives one signal to the cognitive system, the yellow concept is activated, and if it gives another signal, the blue concept is activated: the concepts differ only in the signal they respond to: they are not decomposable any further. The buck stops, so to speak, at the colour quale: if your internal concept-analysing mechanism tries to pick apart the "yellow" concept, or tries to compare it with the "blue" concept, there comes a stage when no further analysis is possible. The "yellow" and "blue" concepts can only say: "Don't ask us WHY we're different: we just ARE".

So, qualia-concepts like "yellow" have this unique property of being unanalysable - but at the same time they are concepts like any other, in that they can be used to build models of the world. In this sense, they are "real" parts of the world to us. It doesn't occur to us that we are mixing two different types of concept: those, like colour concepts, that refer to properties of our sensory input *channels* and those that refer to *patterns of signals coming down those channels.* The two sorts of concept seem in every way compatible, and so we are tempted to ask what colour qualia "are" in the same way that we ask, scientifically, what other things in the world "are".

What is being suggested here is really just an outline sketch of an argument that will need

particular knowledge of the way our cognitive systems work before it can become a convincingly complete resolution of the problem of qualia. The method, though, is clear enough. The structure of cognitive systems may be such that certain concepts that they necessarily build (which refer to, or represent, aspects of their own internal structure) do not correspond to things in the external world, but yet at the same time cannot, at first blush, be distinguished by the system as ficticious or unreal. In consequence, the system can frame questions about these entities, such as "I wonder if everybody else has the same sensation of red when they look at a "red" object?" and "What would it be like to see a new colour?", which cannot be answered, because the entities in question have no external, publicly verifiable existence.

The cognitive system can get itself into much deeper trouble than this. In considering its own nature, for instance, it will have to conclude that science leaves out any mention of the subjective aspects of its world when accounting for the way that it [the cognitive mechanism] functions. It can understand its own mechanisms in an objective way, but it can see no way that qualia can be studied, compared or accounted for. The problem is not restricted to sensory qualia, either. The sense of "self" and the notion of "consciousness" (stripped, as before, of all its contingent features like acuity of cognitive functioning) are similarly unresolvable. Reflecting on the nature of her own identity and existence, even the most hardened materialist, if she is honest, must admit that something *seems* to be left out when she pictures herself as just a machine; as just a myriad of little signals chasing one another down wires (or axons).

This line of argument assumes a materialist account of the nature of the mind, but there is no circularity here: the proposal is that by making this assumption the outstanding questions answer themselves.

It is worth noting that if this is the proper way to tackle qualia, then it leaves them at one and the same time explained and not explained. The subjective world of any given individual is left entirely unaccounted for in objective terms: the only way we can know what it is like to be somebody else is to assume that they are made the same way we are, and suppose that they must experience things like we do, only with perhaps a different amount of emphasis here and there. The less like us they are (in physical or computational structure) the less excuse we have for making the extrapolation.

But while we may not be able to say anything useful about qualia, we may nevertheless be able to account for why we can say nothing about them. If the structure of cognitive systems in general (or even just our own particular sort of system) are understood to give rise to these fictitious concepts, then we would have a clear, objective account of why beings such as ourselves should find certain aspects of the world entirely mysterious.

6) Conclusion

It is an interesting exercise to review the previously discussed alternative positions on reducing mental states to brain states, in the light of this new account. There is a sense in which the epiphenomenalists were right after all. The qualia are a non-physical property of the brain's functioning if by "non-physical" you mean "forever beyond the scope of sciences like physics, neurophysiology and computer science" (Churchland, 1984, p.7). Science can say nothing about your experience of the colour purple, as it looks to you.

Eliminative materialism, too, might be said to have a share in this answer, since the qualia are, in the end, figments of our imagination that will play no significant role in the functioning of cognitive systems at large. This is only a qualified concession to eliminative materialism, however, since the claims it makes are broader, encompassing more than just the qualitative aspects of mental states (and about these claims the qualia argument given here is agnostic).

Finally, orthodox functionalism is perhaps the closest doctrine to the one suggested here: modulo some further details of the mechanisms that might give rise to ficticious concepts, it seems that

the notion of inverted qualia is not coherent (there is no basis for comparison of qualia if they are ineliminably subjective), and absent qualia in an organism that is functionally equivalent to one that does have qualia is not possible. Suggestions by the physicalist-functionalists that the substrate material out of which the cognitive system is built might be relevant to the issue are seen, on this account, to be misguided.

References

Block, N. (1978). "Troubles with functionalism" in: Savage, C.W. (Ed.) *Minnesota Studies in the Philosophy of Science, vol. 9,* University of Minnesota Press.

Block, N. (1980). "Are absent qualia impossible?" *The Philosophical Review, 89,* 257-274.

Churchland, P.M. (1984). "Matter and Consciousness" MIT Press/ Bradford Books.

Harman, G. (1982). "Conceptual role semantics" *Notre Dame Journal of Formal Logic, 23,* 242-256.

Nagel, T. (1974). "What is it like to be a bat?" *The Philosophical Review, 83,* 435-450.

Searle, J. (1980). "Minds, brains and programs" *The Behavioural and Brain Sciences, 3,* 417-457

Shoemaker, S. (1975). "Functionalism and qualia" *Philosophical Studies, 27,* 291-315.

Shoemaker, S. (1981). "Absent qualia are impossible - A reply to Block" *The Philosophical Review, 90,* 581-599.

Shoemaker, S. (1982). "The inverted spectrum" *Journal of Philosophy, 79,* 357-381.

White, S.L. (1986). "Curse of the qualia". *Synthese, 68,* 333-368.