The Knowledge Argument

Frank Jackson

I

The physical sciences tell us a great deal about what our world is like. They also tell us a great deal about what we are like. They tell us, for example, that our bodies are made up of the stuff that the physical sciences — physics, chemistry and biology — talk about. We can think of this as our physical nature, our nature as revealed by the physical sciences, or maybe by certain future developments of the physical sciences. A perennial question is whether our physical nature is our total nature. Is it the case that the physical account of us captures what we are like without remainder, or is there something more to us and, in particular, are our minds or aspects of our minds that something more? This is one way of asking the dualism versus materialism question. Dualism says yes, there is more to us than our physical nature; materialism says that's all there is. Or, more precisely, the kind of materialism we will be concerned with says that's all there is.

There is a weak kind of materialism which holds that each and every part of us is material — there is, for example, no soul as traditionally conceived—but grants that we have special properties different in kind from those inventoried in the physical sciences, and different in kind from those we can think of as constructions out of properties so inventoried. Weak materialism denies rather than affirms that our nature is exhausted by our physical nature and is really a kind of dualism, a dual attribute kind of dualism. Materialism, as we are understanding it, is the real McCoy and is often called physicalism when it is important to keep this in mind.

In the first issue of this journal, Alan Thomas, 'Is Your Mind Your Brain?', canvasses some of the arguments for and against the view that our physical nature exhausts our nature. We will focus on an especially thorny part of the debate between dualists and materialists. It concerns the 'feely' side of psychology, the mental states with a phenomenology, the mental states for which there is something it is like to be in them. These phrasings are different ways of getting at the same general idea: the idea that there is a feel to states like pain, itches, experiences of colour, feelings of heat and so on that is missing in the case of mental states like the belief that oxygen is essential to life or the desire that it will rain soon. Such beliefs and desires may be accompanied by various feelings but in themselves are not feelings and lack any distinctive phenomenology.

Physicalism has a special problem with the mental states with a phenomenology. We often think of cognitive states like belief and conative states like desire in functional terms. Belief is a state typically induced in us by the environment, which carries putative information about the environment, and desire is a state that works with belief qua informational state to make our bodies move in such a way that the environment is changed in various ways—the ways that would satisfy our desires in cases where our beliefs are true. This is, of course, far too crude an account of belief and desire but one gets a glimmering of how something like it might work, and if something like it could be made to work, there would be no threat to physicalism from belief and desire. Functional states are part and parcel of the materialist cum

physicalist conception of what our world is like. The situation is very different, it would seem, in the case of itches, heard sounds, sensings of blue and the like. They appear to have an intrinsic feel connected with our consciousness of them that is left out of account by any functional story. A way to bring out the contrast between belief per se and perceptual experience is to think of what happens when you shut your eyes. You will likely retain some sort of belief about the location, size and colour of the objects around you. But, it seems, 'something' goes when you shut your eyes. What goes is the phenomenological aspect of perceptual experience.

The knowledge argument is one way of seeking to turn the intuition in play in the remarks immediately above into an argument from premises even confirmed materialists find it hard to deny. The argument has a number of forms and has been advanced by many writers. Some references are given at the end. From my possibly biased perspective, I like the following version.

We suppose that we have a brilliant physical scientist, Mary, who is confined in a black and white room. There are no windows. She herself is, we may suppose, painted white all over and dressed in black. All her information about the world and its workings comes from black and white sources like books without coloured pictures and black and white television. However, the lectures she receives over the black and white television and the books she reads are amazing feats of exposition in physics, chemistry, biology and cognitive science, and she has extraordinary powers of comprehension and retention. In consequence, she is, despite the artificial restrictions in which she works, extraordinarily knowledgeable about the physical nature of our world, the neurophysiology of human beings and sentient creatures in general, and how their neurophysiology underpins their interactions with their surroundings including for instance the fact that on many occasions they produce words like 'red' and 'yellow' (if they speak English) when in front of blood and buttercups, respectively.

Can she in principle deduce from all this physical information what it is like to see, say, red? It seems that she cannot. Despite her vast knowledge of the physical facts, there is something about our world and especially about persons' colour experiences she is ignorant of. This conclusion is reinforced by reflecting on what would happen should she be released from her room. Assuming that there is nothing wrong with her colour vision despite its lack of exercise during her imprisonment, she would learn what it is like to see red, and it is plausible that this would be learning something about the nature of our world, including especially the nature of colour experiences. Surely, runs the argument, she could not have predicted this in advance, and surely she would come to realise that her conception of the mental lives of others had been seriously impoverished. It follows that she did not know while in the room all there was to know about our world. But ex hypothesi she did know all there was to know physically. Therefore, there is more to know than all there is to know physically. Physicalism is false.

This argument has attracted some strong supporters (but ones who have typically sought to make one or another improvement to the argument, as is the way of philosophers), and some strong critics. As you would expect given the current and understandable presumption in favour of materialist views of mind, the second group has been larger than the first. The criticisms have been very various; so various as to

constitute close to an empirical refutation of the idea sometimes floated that it is obvious where the knowledge argument goes wrong.

The objections can be usefully categorised in terms of which of the two main claims in the argument is targeted. One claim in the knowledge argument is that complete physical knowledge is not complete knowledge tout court (or anyway not as far as the mind is concerned). The other claim is that if physicalism were true, it would be. From these two claims it follows that physicalism is false by Modus Tollens[1]. Let's call these two claims the incompleteness claim and the deduction claim, respectively. The incompleteness claim is supported by the plausibility of the contention that Mary would learn something on her release. The deduction claim is that were physicalism true, there would be nothing Mary could not in principle work out about what our world is like.

II

Let's look at some objections to the deduction claim. Critics of this claim urge that it is consistent to hold a) that complete physical knowledge is incomplete knowledge of our world, with b) that the physical account of what our world is like is complete.

Sometimes the critics spell this out by pointing out that no amount of knowledge of what one's world is like amounts to knowledge of, for example, who or where one is in one's world. Suppose that some high-powered physics demonstrates that our world will go through two exactly similar cycles and that we know this. In that case we could not possibly know which cycle we were in. There would be no way to tell the difference between being in the first cycle and being in the second cycle. Our surroundings and our bodies, for example, would be indistinguishable whether we were in the first or the second cycle. In particular, Mary would not know which one she was in. She would know that there were two exactly alike people called 'Mary', each living in exactly alike black and white rooms, but would not know which one she was. The moral is that complete knowledge of what our world is like does not necessarily deliver knowledge of where or who one is. A special case will be where complete physical knowledge fails to deliver knowledge of where or who one is. So the idea that physicalism is committed to complete knowledge of the physical delivering complete knowledge simpliciter is a mistake.

This point about the inadequacy of knowledge of what our world is like to deliver knowledge of who and where we are is correct and important. But it is far from clear that it goes to the heart of the knowledge argument. Mary's lack of knowledge seems at least in large part to concern what her world is like, not her or anyone's identity or location in it. She is ignorant before her release, it seems, of what certain experiences are like. That is the point of the argument.

The other main line of attack on the deduction claim starts from the point that the very same things, facts, events and so on can be known under many guises. The FBI may know Jones under the guise of the main suspect in a mail fraud; you may know him as your next door neighbour; I may know him as the person who has just bought an expensive car at my dealership. The same happening may be Jones's arrest; the disturbance next door; the event that means the car is never paid off. This suggests that we could grant that Mary's knowledge of what her world is like is incomplete

without being forced to the conclusion that what is not known is non-physical. Her ignorance is a matter of there being features or categorisations of certain happenings in the world, especially those involving colour experience, that elude her while she is inside the room. All the same, the happenings in question are purely physical ones. When Mary leaves the room, she acquires knowledge but entirely through knowing about the very same physical things, facts, events and so on under different guises or under different categorisations. She gets new ways of categorising happenings around her and thereby acquires new knowledge, but it is, all the same, knowledge of the purely physical and so no threat to physicalism. Consider, for example, someone knows the Cartesian coordinates of a series of points that all lie on a circle without realising that the points lie on a circle. It is not until they graph the points, or do the calculations that reveal that the points satisfy the relevant equation, that they make the discovery. They will learn something, but it is plausible that they do not learn about a new feature—the circularity was 'there' in what they knew from the beginning—the learning was a matter of its becoming revealed when they saw that the points could be categorised in a certain way.

The interest of this suggestion is clear but again it seems that the knowledge argument survives. For the guises, ways of categorising, must all be consistent with physicalism if physicalism is true. But then, it seems, Mary could know about them when inside the room. It is hard to see how given physicalism, there could be ways of grouping things into categories that are, in principle, unavailable to her while in the room. Of course, the ways of grouping may not be easy ones to latch onto. It is easy to miss the fact that a series of points lie on a circle and it can be much harder in more complex cases. But it should be possible in principle to spot the relevant groupings if only one is smart enough and can put the data together aright. However, no amount of cleverness in assembling data and spotting patterns will in itself tell Mary in the black and white room what it is like to see red, or so it seems.

III

Attacks on the incompleteness claim, the claim that as Mary learns (would learn) something new about what the world is like when she leaves the black and white room and so that her knowledge while in the room is incomplete, fall into two broad categories. The first can be introduced by reference to the example of hard to spot patterns that we have just been discussing. The difference between, on the one hand, being in a situation where patterns in, or ways of classifying, data are very hard to grasp although available in principle and, on the other hand, being in one where it is impossible to make the classifications is not always transparent. In consequence, we should insist that Mary can know what it is like to see red while in the room. The strong intuition to the contrary is the result of the fact that it would be extremely hard for her to spot the relevant patterns, along with wrongly conflating the extremely hard with the impossible. This reply is sometimes coupled with the view that the way our brains enable us to see colour goes via the way our brains and optical systems pick up on very unobvious patterns in the effects coloured objects have on light.

Mary's practical problem inside the room will be that her pattern detector, her optical system, is not being allowed to do its job of making sense of data that looks, on the face of it, a complete mess. But if we suppose, as we should when evaluating the knowledge argument, that Mary has worked out everything she could know in

principle (though not in practice) from her vast data bank of physical information, then, runs this reply, we give her knowledge of what it is like to see red. She will not learn what it is like to see red on her release; she will already know.

The second line of attack on the incompleteness claim offers a different diagnosis of the appearance that she acquires knowledge on her release. Instead of the explanation being a slide from the extremely hard to the impossible, the explanation is that we confuse knowledge how with knowledge that. Mary gains knowledge how, not knowledge that.

'Knowledge that' is propositional knowledge, knowledge of the kind of world we live in, of how things are. 'Knowledge how' is knowledge of how to do something: ride a bike, capture a likeness in oils, or recognise a dog that is about to attack. It is an ability. Exercising and acquiring an ability may well require propositional knowledge. Painting classes help us acquire and exercise the ability to capture a likeness in oils, and a lot of propositional knowledge is imparted at these classes: the right paints to buy, good ways of getting skin colour and relative proportions right, and so on. But no amount of propositional knowledge is in itself enough to enable one to capture a likeness in oils. If the knowledge Mary acquires on her release is knowledge how, abilities, there is no problem for physicalism in the knowledge argument. The argument will merely have demonstrated her lack of certain abilities while in the room, not a gap in her knowledge concerning the nature of our world. It will not show that there are features of our world she knows nothing of when in the room despite knowing everything physical there is to know; it will only show that there are things she cannot do.

What abilities does Mary acquire on leaving the room? The usual suggestion is that she acquires the ability to summon up in memory what it is like to see red, to imagine how well a new colour scheme will go, to recognise colours (as opposed to having to ask someone else what something's colour is) and the like.

Everyone agrees that this is part of what happens, would happen, to Mary on her release but there is a persistent intuition that in addition she learns more about how things are. Isn't part of the explanation of her new abilities the fact that she has more knowledge that? Imagining seeing a rhomboid is greatly helped by knowing what a rhomboid is. In the same way, it seems that Mary's new abilities will rest in part on her new knowledge of what it is that she's exercising her abilities on.

Conclusion

The overall situation with the knowledge argument seems to be that the objections to it all make important points but somehow leave one unsatisfied. However the reasons that favour physicalism are very strong. Many feel that the case for physicalism is so strong that one or more of the objections to the knowledge argument must be right, and that the task before us is to find a way of putting the successful objection or objections that removes the feeling of dissatisfaction, or maybe to find an explanation of why there will always be a feeling of dissatisfaction which allows us to discount the significance of the feeling. The latter would be an explaining away of why we are in the grip of the argument.

Frank Jackson
Australian National University

Further Reading

There have been many statements of the knowledge argument, or of arguments close to the knowledge argument in one way or another. The statement above is closest to those in Frank Jackson, 'Epiphenomenal Qualia', and 'What Mary didn't Know'. A recent reprinting of 'Epiphenomenal Qualia' is in David Chalmers, ed., *Philosophy of Mind: Selected Classic and Contemporary Readings* (Oxford: OUP, 2002). Recent reprintings of 'What Mary Didn't Know' are in Frank Jackson ed., *Consciousness* (Dartmouth: Ashgate, 1998), and John Perry and Michael Bratman, eds, *Introduction to Philosophy: Classical and Contemporary Readings*, third edition (Oxford: OUP, 1999). These collections contain many articles discussing the knowledge argument and related matters.

[1] If P then Q, not Q, therefore not P. Put more simply, if physicalism is true (P), then physical knowledge provides a complete account of knowledge (Q). But physical knowledge doesn't provide a complete account of knowledge (not Q), therefore physicalism is not true (not P).