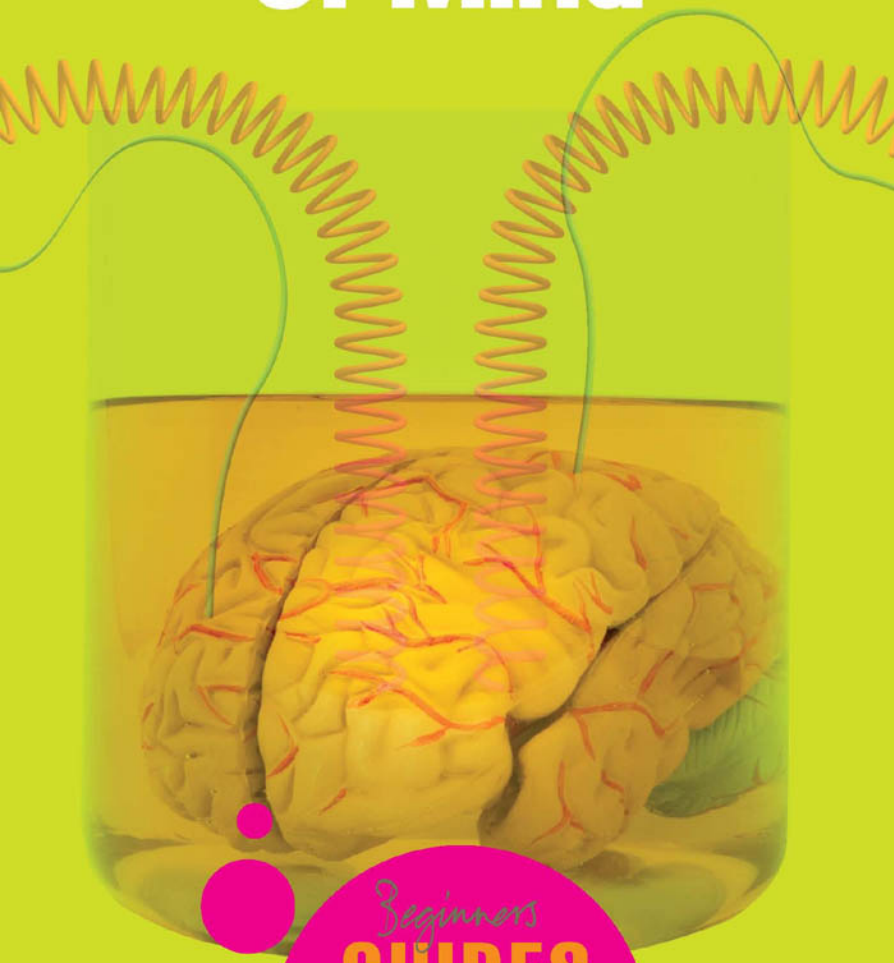
Edward Feser

Philosophy of Mind



Beginners GUIDES

PHILOSOPHY OF MIND: A BEGINNER'S GUIDE

A Oneworld Book
Published by Oneworld Publications 2005
Reissued 2006

A CIP record for this title is available
from the British Library

ISBN-13: 978–1–85168–478–6
ISBN-10: 1–85168–478–6

Typeset by Jayvee, Trivandrum, India
Cover design by Two Associates
Printed and bound in Great Britain
by Biddles Ltd., King's Lynn

Oneworld Publications
185 Banbury Road
Oxford OX2 7AR
England
www.oneworld-publications.com

NL08

The knowledge argument

The zombie argument tries to show that physical reality does not, on its own, add up to mental reality. A related argument, which reinforces this basic idea, tries to show that knowledge of physical reality does not on its own add up to knowledge of mental reality. It is accordingly generally known as the *knowledge argument*, and derives from the contemporary philosopher Frank Jackson.

Jackson asks us to consider Mary, a neuroscientist living in the far future when we have a complete knowledge of the details of the structure and functioning of the nervous system. Mary is in the unique situation of having lived her entire life in a black-and-white room, interacting with the outside world via a black-and-white television monitor. So she has never had any experience of color. (We can even imagine that she has always worn a suit that covers her entire body, and which has kept her from seeing the color of her skin and hair, etc.) While in this room she has come to master the science of the brain, and in particular she has acquired a thorough knowledge of the physics and physiology of color perception. She has never seen the color red herself, but she knows exactly what happens in the eyes, nervous system, and on the surface of the object whenever anyone does see red. She knows down to the last detail, that is to say, all the physical facts there are to know about the perception of color. Now let's imagine that one day Mary is allowed to leave the room, and upon her release she is shown a red apple in full living color for the very first time. Will she learn anything from this experience? Surely she will: she will learn what it is like to see red. And what this shows, according to the argument, is that materialism is false.

The reasoning is this. Materialism claims that the physical facts about perception and the like are all the facts there are. But Mary, hypothetically, knew all the physical facts there were to know about perception – the sorts of facts that could be written down in neuroscience textbooks or conveyed in lectures heard over the

television monitor. Yet she did not know all the facts there were to know about perception, because she learned something new about it upon leaving the room – and you can't learn something you knew already. So what she learned must be a *non*-physical fact. In particular, knowledge about qualia – about what it's like to see red, for instance – must be knowledge about something non-physical.

The suggestion that knowledge of all the relevant physical facts cannot yield knowledge of all the facts about conscious experience has also been illustrated vividly in an example given by Thomas Nagel. Bats, Nagel notes, navigate via senses very different from our own: where we rely chiefly on vision and hearing, they use a kind of sonar or echolocation, putting together a sensory map of the external world by emitting shrieks and then registering the sound waves that bounce back to them from the objects in their immediate environment. The experiences bats have in perceiving the world in this way must be radically dissimilar to ours. Scientific investigation into the structure and functioning of a bat's nervous system may well give us insight into the mechanics underlying its perceptions. But the nature of the perceptual experiences themselves – what it is like to *be* a bat – cannot be revealed by such inquiry, Nagel argues. For science gives us only the objective, third-person facts about any phenomenon, leaving aside any aspect tied to a particular point of view. But it is only from the particular, subjective point of view of a bat that a bat's experiences can be understood. Materialistic scientific accounts must necessarily be inadequate to capture all the facts about a bat's consciousness – or any consciousness, for that matter.

One response sometimes made to arguments like this is that they simply assume that future neuroscience won't be able to explain all there is to explain about conscious experiences: how can we know for sure that Mary wouldn't know what it is like to see red, simply from having mastered the material in her textbooks while in the black-and-white room? There are two problems with

this suggestion. The first is that it seems intuitively implausible. Any facts the neuroscientists of the future are likely to discover are bound to be facts of the same general sort they already know: facts about how neurons are wired, or about which biochemical substances are involved in which processes. It is hard to see how any further knowledge of that sort – of yet more objective, third-person phenomena – could reveal the subjective, first-person facts about what it is like to experience red or to get about by echolocation; there is just a basic and straightforward conceptual difference between the former sort of fact and the latter. The second problem is that the suggestion at hand seems inevitably beset by the same indeterminacy that plagues some versions of physicalism, as we saw in the previous chapter: what if the way neuroscientists of the future explain conscious experience is by positing non-physical properties? This would vindicate the knowledge argument rather than undermine it. Yet there is nothing about the current course of neuroscience that can reasonably lead us to expect any other way in which it might explain consciousness.

More formidable responses to the knowledge argument usually proceed by conceding that there is a sense in which Mary would learn something upon leaving the room, even though she's mastered the neuroscience of the future. The strategy is then to argue that what she learns can, when rightly understood, be seen not genuinely to threaten materialism. Paul Churchland argues that on leaving the room, Mary would not actually learn any new *facts*; rather, she would just learn, in a new *way*, facts she *already* knew. So since she already knew all the physical facts, and there are no new facts (non-physical or otherwise) she learns after leaving the room, the conclusion that the physical facts cannot be all the facts there are is blocked. Churchland elaborates upon this suggestion by appealing to Russell's famous distinction between "knowledge by acquaintance" and "knowledge by description": you might now know about giraffes only by descriptions you've heard or read in a book, but you might someday know about them by becoming

directly acquainted with them in perceptual experience; similarly, Mary, while still in the room, knew all the facts about the experience of red only by description, and then becomes directly acquainted with those very same facts after leaving the room.

One possible objection to this argument is that it seems implausible to suggest that Mary doesn't learn a new fact on leaving the room: surely the fact that red looks like *this* (where "this" refers to the immediate sensation she has of the color) is a fact she did not know before leaving the room, but learns afterward. Another problem is that the Russellian distinction Churchland appeals to is not as philosophically neutral as it might appear. Russell himself held that all we really know by acquaintance are, not external physical objects like giraffes, but rather (what philosophers these days would call) the subjective qualia we normally suppose to have been produced by such external objects; the external physical world in its totality is something we know only indirectly, by description.This goes hand in hand with the sort of indirect realist theory of perception discussed in chapter 1, of which Russell was a proponent (as is Jackson, for that matter). It also raises the question of precisely what these qualia are with which we are directly acquainted; Jackson and (as we'll see in the next chapter) Russell take them to be irreducible to the sorts of properties revealed by physical science, properties which, unlike qualia, we cannot know by acquaintance. So to appeal to Russell's conception of knowledge by acquaintance can hardly help Churchland in rebutting an argument against materialism. But to reject Russell's conception and insist instead that knowledge by acquaintance does not involve knowledge of non-physical qualia would be to beg the question. Either way, it seems that Churchland's response to Jackson's argument fails.

Another response is put forward by David Lewis, who, like Churchland, denies that what Mary learns is a fact she didn't know before. Rather, the knowledge she gets is knowledge of new *abilities*: knowledge of *how* to do something rather than knowledge

that something is the case, and in particular knowledge of how to recognize red objects, the ability to imagine red, and so forth. But this reply seems to have problems parallel to those undermining Churchland's: for one thing, it seems implausible to assert that Mary learns no new facts, since knowledge that red looks like *this* (referring to a subjective sensation) is knowledge of a new fact; for another, the distinction Lewis appeals to is itself not necessarily a neutral one. Mary may well gain new abilities or knowledge upon leaving the room, but it is arguable that some of those abilities are gained only because she learns new facts: Mary now has the ability to imagine what red looks like, but only because she has also learned the fact that red looks like *this*.

Robert van Gulick presents a somewhat technical reply to Jackson's argument. He claims that what Mary gains is knowledge of a new concept, and that if she also learns new propositions this is so only on a fine-grained scheme of individuating or distinguishing between propositions. What this means can best be explained by example. Whether the proposition that water freezes at 32 degrees Fahrenheit and the proposition that H_2O freezes at 32 degrees Fahrenheit are the *same* proposition depends on whether we individuate propositions in a fine- or coarse-grained mode. A fine-grained mode would be one which took account of the fact that "water" and "H_2O" are associated with different concepts (even though they refer to the same substance) and thus would count these propositions as distinct; a coarse-grained mode would ignore the difference in concepts and (since "water" and "H_2O" refer to the same substance) count them as identical. Similarly, the proposition that $5 + 7 = 12$ and the proposition that 38 is the square root of 1,444 are the same proposition on a coarse-grained mode of individuating propositions (one that takes account only of the fact that these mathematical propositions, being necessarily true, both have exactly the same truth value in every possible world); but they are different propositions on a fine-grained scheme, one that takes account of the different concepts

associated with "5," "+," "7," "=," "12," "38," "square root," and "1,444." In the first example, it is clear that even if we count the propositions as different, the fact they refer to is the same: water is identical to H_2O, so the fact that water freezes at 32 degrees Fahrenheit is the same fact as the fact that H_2O freezes at 32 degrees Fahrenheit. Similarly, van Gulick suggests, even if Mary, having learned a new concept after leaving the room, is thereby also able to learn a new proposition, it would not follow that the fact that proposition describes is a fact she didn't already know. Perhaps it is a physical fact of the same sort she already knew while still in the room.

As with the other responses to the knowledge argument, one could object to this one that it seems intuitively implausible: the fact that red looks like *this* (where "this" refers to an immediate sensation) seems obviously to be a different fact than the fact that Mary is in a brain state of type B (or whatever). Of course, van Gulick might suggest that the way things seem might nevertheless in this case be wrong: it might also seem to someone ignorant of chemistry that the fact that water freezes at 32 degrees Fahrenheit is a different fact from the fact that H_2O freezes at 32 degrees Fahrenheit, even though they are in reality the same. But it isn't clear that this suggestion will work. After all, few people would find it a satisfactory defense of the highly dubious claim that the fact that $5 + 7 = 12$ is the same fact as the fact that 38 is the square root of 1,444. In the case of this mathematical example, we surely have two different facts, not just two different fine-grained propositions. Indeed, it is partly our sense that this is so that leads us to see the need for a fine-grained mode of individuating propositions in the first place: we don't suppose this is necessary merely in order to take account of differences in concepts, but also because the propositions of which concepts are constituents often seem (as in the mathematical example) to be about different facts. But the suggestion that the facts that Mary learns on leaving the room are the very same facts as those she knew before seems just as intuitively

implausible as the suggestion that the mathematical facts in our example are the same. And if such an implausibility is, in the one case, itself precisely what leads us to accept a more fine-grained account of mathematical propositions – so that it would be absurd to suppose that one could defend the claim that the mathematical facts in question are the same by appealing to a fine-grained account – then it would be (equally) absurd and implausible to suppose that one could refute the knowledge argument by a parallel appeal to a fine-grained scheme of individuating propositions. In other words, it is in part precisely because it seems so intuitively plausible that facts about qualia and physical facts are just different sorts of fact that we find a fine-grained mode of individuating propositions about them to be necessary in the first place. So it won't do to appeal to such a mode in order to defend the claim that they aren't different.

Subjectivity

Most of the criticisms of the knowledge argument are more or less along the same lines, and would therefore be open to similar objections. But there is another possible reply, suggested by what was said earlier about the inverted spectrum scenario, which may be more formidable. Suppose that each color can indeed be given a precise location in color space, and thus analyzed in terms of its relations to every other color. It then seems possible, at least in principle, that one might be able to deduce the nature of one color from its relations to the others. Consider a simple example involving three very close shades of blue, A, B, and C, where A is the lightest, C the darkest, and B intermediate. It is certainly plausible that someone who had only ever experienced A and C would be able to figure out what it would be like to experience B simply by considering its relations to A and C (the relations being "darker than" and "lighter than"). By extension, it may also be plausible to

suggest that someone who had never seen orange could, in principle, determine what it would be like to experience it if he or she had experienced red and yellow: one could deduce the appearance of orange from its being similar to, and intermediate between, these other colors. Why not conclude, then, that someone who had had at least *some* visual experience – of black and white, of gray as intermediate between them, of light and dark – might in principle be capable of deducing what the various colors looked like based on a sufficiently detailed description of their relations? Why not conclude in particular that Mary – who studied the theory of color and the structure of color space – would have been able in principle to deduce what it would be like to experience red while still in the room, so that she would in fact not have learned anything new when leaving it?

This sort of strategy could in theory be extended to all qualia – auditory, tactile, olfactory and gustatory as well as visual – which could all be described in terms of their relations to other qualia of the same sort, and even their relations to qualia of different sorts: "warmth," "coolness," "hardness," softness," "sharpness," smoothness," seem to be qualities applicable to many different kinds of qualia, so that (to some extent at least) visual qualia can be described in terms of their similarity relations to auditory qualia, auditory qualia in terms of their similarity relations to tactile qualia, and so forth. Rudolf Carnap (1891–1970) attempted just such a detailed and systematic analysis of all qualia in terms of their relations to each other, which relations he took to be grounded ultimately in the basic relation of "recollection of similarity." If such an analysis could be carried out completely, then it is arguable that anyone thoroughly familiar with it could, on the basis of even the most limited sensory experience, determine what it would be like to have any experience that he or she has never in fact had.

This approach seems promising, though it would take a great deal of argument convincingly to defend it. But even if successful, the critic of materialism could hold that this strategy would not

undermine the deeper truth captured by the knowledge argument, in Nagel's version more than in Jackson's. That truth is, arguably, just this: while Mary might at least in principle be able to deduce, from what she knows while still in the room, what it is like to experience red, she would not be able to deduce from it *why it is like anything at all*. The real mystery is not that red "feels" specifically like this rather than that; it is that it has any "feel" in the first place. Nagel captures the problem by noting that it is the fact that there is "something it is like" to be conscious that makes consciousness so difficult to account for in purely material terms. The zombie argument captures it by suggesting that it is metaphysically possible for there to be creatures physically identical to us but without consciousness, creatures who exhibit exactly the same behavior – and thus, for example, make exactly the same discriminations between red and other colors – but who do not experience red, for whom there is nothing it is like to discriminate red from other colors. That there is something it is like for us to experience it would seem to be a further fact about us, over and above the physical ones.

This goes hand in hand with Nagel's point that a conscious being is one with a first-person point of view on the world, who is a locus of subjectivity. Consciousness of what an experience is like is always consciousness of what it is like "for me," for a subject of experience; and for Mary to deduce what experiencing red would be like from its similarity relations to other experiences presupposes that she is a conscious subject *for whom* it would be similar. One might think to deflate this notion of subjectivity by suggesting that lots of purely physical things have points of view on the world as well – a camera, for instance, which can photograph only what is in front of it; its images produced by reflecting its particular point of view – so that it shouldn't be so mysterious why we, with our specific sensory organs and physical limitations, should have points of view too. But such a suggestion would seem fallacious. A camera is just a mechanism sensitive to light such that it

can be used to generate patterns on film that correspond to the light patterns reflected by physical objects. It has no literal "point of view," for it doesn't view anything in the first place in the sense in which we do. It is we who understand the pictures the camera produces to have significance – indeed, it is we who regard them as pictures rather than splotches of chemicals on paper. It is also true that the particular point of view any of us occupies is, like the camera, limited by our specific position in space and the physical constraints imposed by the structure of the human body. But (to make a point that parallels the point made above about the experience of seeing red) it is not our having this or that particular point of view that is claimed to be difficult or impossible to explain in materialistic terms; it is rather our having any point of view at all that is mysterious.

In the dualist's view, that science, at least as understood by materialists, cannot *in principle* solve this mystery seems to follow necessarily from the very nature of scientific explanation: it is not a matter of our not yet having gathered all the relevant neurological evidence or hit upon the right theory. For, as noted in the last chapter, the method of modern scientific explanation has historically been precisely to carve off and ignore the subjective, observer-relative aspect of any phenomenon it investigates and identify such phenomena exclusively with the objective, third-person residue which remains. We can take the explanation of temperature as a paradigm. A hoary philosophical example illustrates the subjectivity of temperature considered as a felt experience: someone who first puts his or her right hand in a bucket of ice cold water and his or her left in a bucket of hot, then puts both in a bucket of lukewarm water, will find that the lukewarm water feels warm to the right hand and cold to the left. We can also imagine extraterrestrials who would feel what we would call coolness when putting their hands (or tentacles) in hot water and heat when putting them in ice cold water. If by "heat" and "cold" we mean the subjective sensations or feelings produced by hot and cold objects,

there is no objective fact about whether a particular object is hot or cold. Science thus ignores subjective feelings and instead defines (or re-defines) heat and cold exclusively in terms of the objective, mind-independent physical facts which (in us, anyway) cause the relevant sensations: facts about mean molecular kinetic energy. But if the method of science is in every case to strip away the subjective appearance a phenomenon exhibits and, as it were, push it into the mind, it seems obvious that the same procedure cannot in principle be applied to an explanation of the mind itself: for the mind *just is* (in part) the collection of the subjective appearances of the things it experiences; the subjective element cannot in *this* case be stripped away without thereby stripping away and ignoring the very phenomenon to be explained – in which case it hasn't really been explained at all.

Subjectivity – comprising the phenomena of being present to an experiencing subject, of being directly accessible only from the point of view of that subject, and of being capable of existing in experience even when (as in dreams or hallucinations) an apparent objective correlate of the experience does not exist – thus appears to be the essential core to the concept of qualia, and the feature that is most plausibly inexplicable in physical terms. Philosophers often attribute other supposedly problematic features to qualia, such as ineffability and intrinsicality, but to a very great extent these appear to be reducible to or parasitic upon subjectivity. For example, qualia seem ineffable only because our language is typically used to communicate thoughts about *objective*, *public* phenomena, and words are typically learned by reference to such phenomena; communicating thoughts about private and subjective phenomena thus seems difficult or impossible. To the extent that qualia are ineffable, this is just a consequence of their being subjective.

Qualia are often claimed to be intrinsic in the sense of not being analyzable in terms of their relations to other things, for example, in terms of the causal relations functionalism claims all mental

phenomena can be analyzed in terms of; for, as was suggested by the zombie argument, it seems logically possible for any such set of causal relations to exist without qualia. But here too subjectivity seems to be what's really at issue. It is because qualia are not analyzable into relations instantiated in *objective*, *third-person* phenomena – causal relations between firing patterns in clumps of neurons, say – that they seem to be intrinsic. Yet this leaves open that they may be analyzable into *subjective*, *first-person* similarity relations of the sort Carnap, Clark, and Hardin have tried to elucidate: that they may well in this sense be *both* irreducibly subjective and yet non-intrinsic. Indeed, it is arguable that it is precisely because they are so analyzable that we can communicate about them despite their subjectivity (so that they are *not* ineffable in the strict sense): if we were not able to describe and convey to one another the systematic similarities and differences between qualia, we would not be able to know (as we surely do know) that we are all talking about the *same* phenomena when we discuss qualia and argue about whether materialism can account for them. Our knowledge of the relational structure of qualia makes our claims about them cognitively meaningful and rationally assessable, despite the fact that the relations comprising that structure are directly knowable only from the subjective, first-person point of view.

It seems arguable then that the key difference between qualia on the one hand and such physical phenomena as functional organization, neurophysiology, and behavior on the other, is that the former are irreducibly subjective, "private," and first-person in character while the latter are inherently objective, publicly accessible, and third-person. The dualist concludes that since the two sorts of phenomena have such irreconcilable essential properties, the former cannot be accounted for in terms of the latter – in which case materialism, which claims that everything real is explicable in terms of objective, third-person physical phenomena, must be false.