

For Patrick Wilken, Tim Bayne, and Axel Cleeremans (eds.), *Oxford Companion to Consciousness*

Yujin Nagasawa
Department Of Philosophy
University of Birmingham
Edgbaston
Birmingham B15 2TT
United Kingdom
Y.Nagasawa@bham.ac.uk

Knowledge Argument

The knowledge argument is an argument against physicalism that was first formulated by Frank Jackson in 1982. While Jackson no longer endorses it, it is still regarded as one of the most important arguments in the philosophy of mind.

Physicalism is the metaphysical thesis that, roughly speaking, everything in this world—including tables, galaxies, cheese cakes, cars, atoms, and even our sensations—are ultimately physical. The knowledge argument attempts to undermine this thesis by appealing to the following simple imaginary scenario:

Mary is confined to a black-and-white room, is educated through black-and-white books and through lectures relayed on black-and white television. In this way she learns everything there is to know about the physical nature of the world. She knows all the physical facts about us and our environment, in a wide sense of ‘physical’ which includes everything in *completed* physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this, including of course functional roles. (Jackson 1986, p. 291)

The knowledge argument says that if physicalism is true, Mary knows everything in this world. However, it seems obvious that her knowledge is not yet complete. Suppose that

Mary leaves her black-and-white environment for the first time in her life. She will then apparently *learn* something; namely, *what it is like to see colour*. Given that she knows everything physical in her black-and-white environment and that she still learns something upon her release, it seems reasonable to conclude that physicalism is false.

Jackson presents the knowledge argument schematically as follows:

(1) Mary (before her release) knows everything physical there is to know about other people.

(2) Mary (before her release) does not know everything there is to know about other people (because she *learns* something about them on her release).

Therefore,

(3) There are truths about other people (and herself) that escape the physicalist story. (Jackson 1986, p. 293)

Jackson formulates premisses (1) and (2) in terms of truths about other people in order to emphasise that Mary certainly learns some truths about the external world.

Given the simplicity of the Mary scenario, it is not surprising to find that similar scenarios had already been envisaged by many other philosophers. For instance, the 17th century philosopher John Locke writes as follows:

I think it will be granted easily that if a child were kept in a place where he never saw any other than black and white till he were a man, he would have no more ideas of scarlet or green, than he that from his childhood never tasted an oyster, or a pine-apple, would have of those relishes. (Locke 1689, Book II, Chapter 1, Sect 6.)

Or, to take another example, Paul E. Meehl (1966) discusses a scenario in which two persons, *K1* and *K2* share 'knowledge of the Utopian scientific network'. While the knowledge of *K1* and *K2* includes 'the psychophysiology of vision and the psycholinguistics of color language' only *K2* is congenitally blind. Meehl points out that many people would think intuitively that in this case *K1* knows something that *K2* does not.

Jackson's knowledge argument is impressive, not because of the intuition about Mary's knowledge upon which it is based, but because of its dialectic structure. First, Jackson formulates the argument for the specific purpose of refuting physicalism. Locke does not derive any claim about physicalism from his thought experiment; he merely uses it to illustrate his empiricism. While Meehl does consider whether his example undermines physicalism, he concludes that it does not. The second distinctive characteristic of the knowledge argument is that it is designed in such a way that while its premisses seem to be ontologically neutral, it derives a significant ontological conclusion. Both premisses (1) and (2) of the argument are apparently innocuous claims about Mary's knowledge, which do not, if taken individually, force one to endorse any specific ontological view. However, once one accepts these premisses at the same time one cannot but reject physicalism, which is certainly a significant ontological commitment.

While most philosophers affirm the validity of the knowledge argument, they are divided as to its soundness. Some philosophers accept it and endorse dualism. However, many others reject it. Moreover, even among those who reject it, there is no consensus at all as to exactly what is wrong with it. During the course of the dispute, a number of distinct objections to the argument have been proposed.

As noted earlier, the knowledge argument is based on the intuition that even Mary's complete physical knowledge does not subsume everything in this world. The most straightforward reaction to the argument is to reject this intuition. According to such philosophers as Jeff Foss (1989) and Daniel C. Dennett (1991), it is a mistake to think that Mary learns something upon her release. If we take physicalism seriously, they claim, we can safely conclude that she will not learn anything when she leaves her black-and-white environment.

However, most philosophers share the intuition that upon her release, Mary does learn something, or at least something epistemologically significant happens to her. Yet it is contentious whether that entails the falsity of physicalism. Laurence Nemirow (1990) and David Lewis (1988), for instance, defend physicalism from the knowledge argument by appealing to the distinction between knowledge-that and knowledge-how. According to their 'ability hypothesis', what Mary acquires upon her release is not new knowledge-that, *i.e.*, propositional knowledge, but knowledge-how, *i.e.*, a set of abilities. When she looks at a red tomato, for instance, Mary will acquire various abilities, such as how to identify red objects as red, imagine and remember a red experience, and so on. If Mary did acquire knowledge-that upon her release, then it would seem that physicalism is false. However, according to proponents of the ability hypothesis, the mere fact that she acquires new abilities does not prove the falsity of physicalism. Defenders of the knowledge argument, by contrast, are dissatisfied with the ability hypothesis because even if Nemirow and Lewis are right in saying that Mary does gain new abilities upon her release, it still seems implausible that they are *all* she acquires.

Earl Conee (1994), John Bigelow and Robert Pargetter (1990) defend the ‘acquaintance hypothesis’, which is comparable to the ability hypothesis. This hypothesis relies on the distinction between knowledge by description and knowledge by acquaintance. According to the proponents of the hypothesis, when Mary has a colour experience she gains only knowledge by acquaintance, *i.e.*, she only becomes acquainted with the experience. Given that Mary does not gain any new knowledge by description, *i.e.*, propositional knowledge, they claim, the knowledge argument fails to refute physicalism. The acquaintance hypothesis seems to have the same difficulty that the ability hypothesis has; while Mary does seem to gain knowledge by acquaintance upon her release, it is not clear that that is *all* she gains.

Unlike proponents of the ability and acquaintance hypotheses, some critics accept that Mary does acquire propositional knowledge upon her release. Nevertheless, they think that physicalism can still be defended. For, according to them, what Mary acquires upon her release is not new propositional knowledge but *old* propositional knowledge *in a new mode of presentation*. She merely regains, they claim, knowledge that she has already acquired in her black-and-white environment in a different mode of presentation. In order to illustrate this point, consider the following two sentences:

(i) Superman can fly.

(ii) Clark Kent can fly.

While (i) and (ii) attribute the same property to the same individual, Lois Lane knows only what is expressed as (i), but not (ii). This is because in order to know what is expressed as (ii) she needs to be in a new mode of presentation, which is distinct from the one that corresponds to the knowledge of (i) (Terry Horgan, 1984). Some philosophers

elaborate upon this response further and defend ‘*a posteriori* physicalism’. According to *a posteriori* physicalism, while phenomenal truths are entailed by physical truths, the entailment is not *a priori* but *a posteriori*. It appears that Mary can know what it is like to see colour without having a colour experience if (i) phenomenal truths are entailed by physical truths and (ii) the entailment is *a priori*. *A posteriori* physicalism accepts (i) but rejects (ii). *A posteriori* physicalists think, in other words, that if the knowledge argument undermines anything, it undermines only *a priori* physicalism, which holds both (i) and (ii). This is the most common objection to the knowledge argument. Defenders of the knowledge argument try to undermine this objection by arguing either (a) that if physicalism true, then *a priori* physicalism must be true or (b) that the knowledge argument can be reformulated so that it is directed against *a posteriori* physicalism as well.

All of the objections to the knowledge argument discussed so far assume that Mary does have complete physical knowledge before her release. However, some critics are sceptical about this assumption. Torin Alter (1998), for instance, says that this assumption is tenable only if all physical truths can be learned discursively. That is, the Mary scenario makes sense only if all physical truths can be learned by reading black-and-white books and watching black-and-white television; but it is not obvious that they can be learned in such a limited way. To take another example, Daniel Stoljar (2006) argues as follows. It is reasonable to believe that we are ignorant of some type of physical truths relevant to phenomenal experiences. If we assume that these truths are covered by the quantifier ‘all physical truths’ that occurs in the description of the Mary scenario, we have no reason to think that she learns something upon her release at the same time as

knowing everything physical. On the other hand, if we assume that these truths are not covered by the quantifier, then the knowledge that Mary acquires prior to her release is not *complete* physical knowledge.

In his 1982 paper, Jackson endorses epiphenomenalism. Given that the knowledge argument seems to refute physicalism and that interactionism is implausible, the only reasonable option left for him seems to be epiphenomenalism. However, some philosophers hold that the knowledge argument is *inconsistent* with epiphenomenalism. Epiphenomenalism claims that qualia are causally inefficacious in the physical world. Ironically, this claim appears to contradict the Mary scenario, for if qualia really are causally inefficacious in the physical world, then surely she does not come to know anything by having colour qualia upon her release. Therefore, according to this objection, one cannot consistently accept both the knowledge argument and epiphenomenalism at the same time. While this objection does not show exactly which premiss of the knowledge argument is false it does show, if it shows anything, that there must be something wrong with the argument. This objection is obviously based on a version of the causal theory of knowledge, which itself is a matter of controversy.

Paul M. Churchland (1989) provides an objection to the knowledge argument in the same vein. According to him, there must be something wrong with the knowledge argument because if the argument successfully refuted physicalism it would equally successfully refute some versions of dualism as well. Suppose, for example, that substance dualism is true and that in her black-and-white environment Mary learns not only all truths about the physical entities, but also all truths about mental substance. That is, she learns everything about the causal, relational and functional roles of physical

entities as well as of mental substance. However, it still seems obvious that she learns something when she has a colour experience for the first time. Therefore, Churchland concludes, the knowledge argument is unreasonably strong.

As I noted earlier, Jackson no longer endorses the knowledge argument. In his second postscript published in 1998, he declared that he had come to think the knowledge argument failed to refute physicalism. Moreover, in his 2003 paper, he introduced and explained in detail his own objection to the knowledge argument. In constructing his objection he appeals to representationalism, according to which phenomenal states are representational states. He says that what happens to Mary upon her release is not to learn new nonphysical truths, but merely to be in a new kind of representational state. While this position might appear similar to the new mode of presentation response mentioned above, Jackson characterises it as a version of the ability hypothesis. For, unlike many proponents of the new mode of presentation response, he rejects the idea that Mary acquires any propositional knowledge, whether it is old or new, upon her release. Mary merely comes to be in a new representational state without acquiring or reacquiring any knowledge. Mary acquires instead, according to Jackson, abilities to recognise, imagine and remember the new representational state.

Along with the conceivability argument and the explanatory gap argument, the knowledge argument is regarded as one of the greatest objections to physicalism. While there are a number of strong arguments for physicalism, any version of physicalism that is vulnerable to the knowledge argument is inadequate.

All the materials referred to in this entry, except the following six, are reprinted in:
Ludlow, P., Nagasawa, Y. and Stoljar D., eds. (2004). *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. MIT Press, Cambridge, MA.

Alter, Torin, (1998). A Limited Defence of the Knowledge Argument. *Philosophical Studies*, **90**, 35-56.

Foss, J. (1989). On the Logic of What It Is Like to be a Conscious Subject. *Australasian Journal of Philosophy*, **67**, 205-20.

Locke, John (1689). *An Essay on Human Understanding*.

Meehl, P. E. (1966). The Complete Autocerebroscopist. In P. Feyerabend and G. Maxwell, eds. *Mind, Matter, and Method: Essays in Philosophy and Science in Honor of Herbert Feigl*, 103-180. University of Minnesota Press, Minneapolis.

Nemirow, L. (1990). Physicalism and the Cognitive Role of Acquaintance. In W.G. Lycan, ed. *Mind and Cognition: A Reader*, 490-499. Blackwell, Oxford.

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