Knowing Our Sensations: Jackson's Argument

From chapter 8 of *Neurophilosophy: A Unified Science of the Mind/Brain*
by Patricia Churchland (1986)

Frank Jackson (1982) has constructed the following thought-experiment. Suppose that Mary is a neuroscientist who has lived her entire life in a room carefully controlled to display no colors, but only shades of white, gray, and black. Her information about the outside world is transmitted to her by means of a black-and-white television. Suppose further that one way or another she comes to know everything there is to know about the brain and how it works. That is, she comes to understand a completed neuroscience that, among other things, explains the nature of thinking, feeling, and perception, including the perception of colors. (This is all wildly unlikely, of course, but just suppose.)

Now for the argument: despite her knowing everything there is to know about the brain and about the visual system, there would still be something Mary would not know that her cohorts with a more regular childhood would, namely, the nature of the experience of seeing a red tomato. Granted, she knows all about the neural states at work when someone sees a red tomato—after all, she has the utopian neuroscience at hand. What she would not know is *what it is like to see red*—what it is like to have that specific experience. Conclusion: her utopian neuroscience leaves something out. This omission implies that there is something in psychology that is not captured by neuroscience, which in turn implies that psychology cannot be reduced to neuroscience.

More formally and with some simplifications, the argument is this:

(F)

(1) Mary knows everything there is to know about brain states and their properties.

(2) It is not the case that Mary knows everything there is to know about sensations and their properties.

Therefore:

(3) Sensations and their properties ≠ brain states and their properties.

The argument is very interesting, and it gives an unusually clean line to the intuition that mental states are essentially private and have an irreducibly phenomenological character. Nonetheless I am not convinced, and I shall try to explain why.

First, I suspect that the intentional fallacy [i.e., the treatment of 'intentional' properties (such as 'being-known-by-me') as though they were genuine properties of objects] … haunts the premises of argument (F). That aside, there are perhaps more revealing criticisms to be made. Paul M. Churchland (1985) and David Lewis (1983) have independently argued that "knows about" is used in different senses in the two premises.

As they see it, one sense involves the manipulation of concepts, as when one knows about electromagnetic radiation and can use the concept "electromagnetic radiation" by having been tutored in the theory. The other sense involves a prelinguistic apprehension, as when one knows about electromagnetic radiation by having had one's retina stimulated in the light of day, though one cannot use the expression "electromagnetic radiation." The latter sense may involve innate dispositions to make certain discriminations, for example. If the first premise uses "knows about" in the first sense and the second uses it in the second sense, then the argument founders on the fallacy of equivocation.

The important point is this: if there are two (at least) modes of knowing about the world, then it is entirely possible that what one knows about via one method is identical to what one knows about via a different method. Pregnancy is something one can know about by acquiring the relevant theory from a medical text or by being pregnant. What a childless obstetrician knows about is the very same process as the process known by a pregnant but untutored woman. They both know about pregnancy. By parity of reasoning, the object of Mary's knowledge when she knows the neurophysiology of seeing red might well be the very same state as the state known by her tomato-picking cohort. Just as the obstetrician does not become pregnant by knowing all about pregnancy, so Mary does not have the sensation of redness by knowing all about the neurophysiology of perceiving and experiencing red. Clearly it is no argument in support of nonidentity to say that Mary's knowledge fails to cause the sensation of redness. Whyever suppose that it should?

There is a further reservation about this argument. With the first premise I take no issue, since we are asked to adopt it simply for the sake of argument. The second premise, in contrast, is supposed to be accepted because it is highly credible or perhaps dead obvious. Now although it does have a first blush plausibility, it is the premise on which the argument stands or falls, and closer scrutiny is required.

On a second look, its obviousness dissolves into contentiousness, because the premise asks me to be confident about something that is too far beyond the limits of what I know and understand. How can I assess what Mary will know and understand if she knows *everything* there is to know about the brain? Everything is a lot, and it means, in all likelihood, that Mary has a radically different and deeper understanding of the brain than anything barely conceivable in our wildest flights of fancy.

One might say well, if Mary knew everything about *existing* neuroscience, she would not know what it was like to experience red, and knowing *absolutely* everything will just be more of the same. That is an assumption to which the property dualist is not entitled to help himself. For to know everything about the brain might well be qualitatively different, and it might be to possess a theory that would permit exactly what the premise says it will not. First, utopian neuroscience will probably look as much like existing neuroscience as modern physics looks like Aristotelian physics. So it will not be just more of the same. Second, all one need imagine is that Mary internalizes the theory in the way an engineer has internalized Newtonian physics, and she routinely makes

introspective judgments about her own states using its concepts and principles. Like the engineer who does not have to make an effort but "sees" the world in a Newtonian manner, we may consider that Mary "sees" her internal world via the utopian neuroscience. Such a neuroscience might even tell her how to be very efficient at internalizing theories. It is, after all, the premise tells us, a *complete* neuroscience.

Intuitions and imaginability are, notoriously, a function of what we believe, and when we are very ignorant, our intuitions will be correspondingly naive. Gedanken-experiments [i.e., thought-experiments] are the stuff of theoretical science, but when their venue is so surpassing distant from established science that the pivotal intuition is not uncontroversially better than its opposite, then their utility in deciding issues is questionable.

Moreover, intuitions opposite to those funding premise (2) are not only readily available, they can even be fleshed out a bit. How can I be reasonably sure that Mary would not know what a red tomato looks like? Here is a test. Present her with her first red object, and see whether she can recognize it as a red object. Given that she is supposed to know absolutely *everything* there is to know about the nervous system, perhaps she could, by introspective use of her utopian neuroscience, tell that she has, say, a gamma state in her O patterns, which she knows from her utopian neuroscience is identical to having a red sensation. Thus, she might recognize redness on that basis.

The telling point is this: whether or not she can recognize redness is clearly an empirical question, and I do not see how in our ignorance we can confidently insist that she must fail. Short of begging the question, there is no a priori reason why this is impossible. For all I know, she might even be able to produce red in her imagination if she knows what brain states are relevant. One cannot be confident that such an exercise of the imagination must be empirically impossible. To insist that our make-believe Mary could not make introspective judgments using her neuroscience *because* mental qualia are not identical to brain states would, obviously, route the argument round in a circle.

How could an alchemist assess what he could and could not know if he knew everything about substances? How could a monk living in the Middle Ages assess what he could and could not know if he knew everything there was to know about biology? He might insist, for example, that even if you knew everything there was to know about biology, you still would not know the nature of the vital spirit. Well, we still do not have a complete biology, but even so we know more than this hypothetical monk thought we could. We know (a) that there is no such thing as vital spirit, and (b) that DNA is the "secret" of life—it is what all living things on the planet share.

The central point of this reply to Jackson has been that he needs independent evidence for premise (2), since it is palpably not self-evident. It cannot be defended on a priori grounds, since its truth is an empirical question, and it cannot be defended on empirical grounds, since given the data so far, as good a case can be made for the negation of premise (2) as for premise (2) itself. I do not see, therefore, how it can be defended.