A response to Mary

Jenann Ismael[a] and Carlo Rovelli[b]

[a] Department of Philosophy, John Hopkins University, Baltimore, MD 21218-2625, USA

[b] CPT, Aix-Marseille Université, Université de Toulon, CNRS, Case 907, F-13288 Marseille, France.

January 2nd, 2025

Abstract

Frank Jackson raised a formidable challenge against physicalism in the form of a fable: Mary comprehends the physics of color vision but has never seen red; when she does, she learns what red looks like. Hence there is knowledge that transcends what is accessible from a purely third-person perspective. We point out that this can be true without contradicting physicalism. The solution of the apparent paradox is to notice that physicalism implies that knowledge must be physically realized. In turn, this implies the existence of (physical) reflexive knowledge, distinct from the knowledge obtained from a third-person perspective.

***

In two celebrated papers, Frank Jackson raised a formidable challenge against physicalism, in the form of a fable.[1] Mary is a scientist that has never seen red but has studied the physics of vision. One day she sees red and learns what red looks like. Jackson characterizes physicalism as demanding that if Mary knows all the physical facts about us and our environment, she knows everything there is to know. He then claims that the fact that Mary learns something new shows that the physics of vision could not teach everything there is to know about red. The three following statements, that is, are incompatible.

(a) Everything that is the case is something physical that is the case (physicalism).

(b) At some time, Mary knows everything physical that is the case regarding color vision (assumption).

(c) Later, Mary gets to know something that is the case regarding her color vision which was not already among the things she knew before (assumption).

The argument assumes that to know what red looks like is not something physical, because without this assumption (b) would imply that Mary "knows what it is like to see red" and (c) would be false. The force of the argument is based on the distinction between third-person knowledge and first-person knowledge. That is, the fable assumes that Mary only knows the full physics of what

[1] Frank Jackson: Epiphenomenal Qualia. *Philosophical Quarterly* 32 (1982) 127, and What Mary Didn't Know, *The Journal of Philosophy* 83 (1986) 291-295. For extensive references, see Nida-Rümelin, Martine and Donnchadh O Conaill, Qualia: The Knowledge Argument *The Stanford Encyclopedia of Philosophy* (Spring 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/spr2024/entries/qualia-knowledge/.

happens in the brain, while she has not had the experience of seeing red herself. So, the proper formulation of the argument is to replace (b) by

(b') At some time, Mary knows everything physical that is the case regarding color vision that she could have learned from textbooks that detail everything that happens in someone's brain when they see red and not experiencing red herself.

What the argument then shows is only that (assuming (c), namely that our intuition about Mary's learning is correct) knowing the full physics of what happens in someone's brain when they see red, that is, from a third-person perspective, does not exhaust everything there is to know about red. As such, the argument is correct: its conclusion is that

(d) There is something more that can be learned, besides everything that can be learned from a third-person perspective.

We see no compelling reason to challenge this conclusion. The interesting question, rather, is whether this conclusion is compatible with physicalism. That is, whether (d) this is compatible with (a). Here we point out that it is.

The alleged incompatibility depends on failing to treat knowledge as itself a physical state. Namely on taking knowledge to be a relation between reality and an abstract space of representation. Namely depends on assuming a violation of physicalism. Of course, there is no surprise that we could prove physicalism wrong by violating physicalism in the hypotheses. If instead we stay faithful to physicalism, Jackson's challenge evaporates because (d) is compatible with (a). To see that (a) can be compatible with (d), we must unpack what "to know" and "to learn" mean in physical terms. We must unpack the physical difference between knowing something from the third-person perspective and knowing something from the first-person perspective.

Let's start from the third-person perspective. For Mary to know, say, where Robert is, something in Mary's brain must be correlated with where Robert is. If Robert is in China, there is some physical arrangement or some process in Mary's brain; if Robert is in India, there is something different. This correlation may not exhaust what it is for Mary to know Robert's whereabouts, but it is certainly a necessary condition for her to know. Let's consider all conditions of her knowledge and let's call K-states all physical states of Mary which we characterize as her "knowing where Robert is". Similarly, let's call K-states, the states of Mary's brain which we characterize as she "knowing what happens in somebody's brain when this person sees red".

Now let's consider the first-person perspective. When Mary sees red, she ends up into a physical state that we characterize as "Mary knows what red looks like". This —assuming physicalism is true[2]— is a physical state. Let's call any such state an R-state. Jackson's acute observation is that an R-state may not be a K-state.

But is this surprising? It is incompatible with (a)? Imagine we program a computer to analyze the behavior of other computers, to check if they are broken or not. This is common technology nowadays. A good computer programmed in this manner can be said to have knowledge of

[2] In this paper we are not arguing for physicalism: we are only showing that Jackson's argument is not a challenge against it.

whether another computer is broken or not.[3] This is for the computer to be in a K-state regarding the knowledge of what is it for (another) computer to be broken. Now suppose the computer we have so programmed itself breaks. Call its resulting state an R-state. There is clearly no reason to think that an R-state must be a K-state for the computer: to know that something is broken is physically something else, for the computer, from being broken or having been broken.

Jackson's observation is thus that an R-state may not be a K-state, which is a physical formalization of (d). Is this a challenge to physicalism, namely to (a)? It is not. Precisely as shown above, this conclusion is compatible with a fully physical account of facts, namely consistent with (a).

Why then does Jackson's argument sound so compelling? The reason is that when we think about "knowledge" we often do not think of it as an embodied physical fact, but we rather see it as living outside physics, in an abstract realm. If we take this step (that violates physicalism) Jackson's argument suddenly bites, because if knowledge resides outside the physical reality, then its object must be the totality of physical facts and there is no more room for a genuine first-person perspective understood as physically embodied (Mary's physical R-state).

If instead we keep in mind that knowledge is embodied, then the subject of knowledge is necessarily a physical system (which is physically correlated with the object of knowledge). But if the subject of knowledge is a physical system, then reflexive knowledge, knowledge about the subject of knowledge is intrinsically different from knowledge about something else.

The key point is that to know things about X is different from being X. And to know you must be something. For a computer to have all possible information about what it is to be broken is different from being broken and it is different from the knowledge of being (or having been) broken. There is, in other words, a kind of knowledge that is first-personal. The existence of such knowledge does not contradict physicalism because first-person knowledge is not unphysical: it is just a physical state that one can be in. Therefore, I can know everything there is to know about the state of seeing red, without being in that state or having ever been in that state and therefore not knowing what it is for me to experience red. What is it for me to have experienced red is precisely to be in the physical state in which I am after having experienced red. Which is something very different from being in the state in which I have information about the physical states of others that have experienced red. All this can be described in physical terms, hence (d) is compatible with (a).

The intuition shared by many people that knowing what it is for me to experience red is different from the knowledge of what happens physically in others when they see red is therefore exactly correct. But it is not an intuition that undermines physicalism, because as soon as we treat knowledge physically, we realize that physicalism not only accommodates but actually *requires* the distinction between third-person knowledge and first-person knowledge. It is physicalism itself that implies that there is an essentially first-personal way of knowing; a special kind of knowledge that only I can have, of states by being in them, and only have of myself by being myself.

Such reflexive knowledge is a state of knowledge, is physically embodied, and takes oneself or one's own states as objects. We know about ourselves; we remember our experiences. If physicalism is

[3] If you think other conditions have to be in place for knowledge attributions to be satisfied, imagine those conditions are in place. Insofar as those are purely physical conditions there's no reason (relevant to the present argument) that they (or relevant analogues) shouldn't be reproducible in a computer.

true, our experiences are themselves physical states that we are in. To have a memory of these experiences is equally to be in a physical state.

Reflexive knowledge is typically expressed with indexicals. The structure of states that express reflexive knowledge is different from the structure of states that express third-person knowledge. States that express third person knowledge typically contain a term that represents something in the world in a way that doesn't depend on who utters them or the state the person is in. States of reflexive knowledge contain terms like `I', or 'this state', that pick out the person uttering the words or the state the person is in. There is nothing mysterious about states of reflexive knowledge: as soon as we realize that any state of knowledge is itself a physical state, it is not hard to see how the states themselves or the systems whose states they are can be referred to in this way.

The fact that Jackson assumes that knowledge is nonphysical, therefore, implies that he assumes that physicalism is wrong. This is evident from his remark that "physicalists must hold that complete physical knowledge is complete knowledge simpliciter." If knowledge is physical, it is only a relation between physical systems, it is always held by some system, and it can be complete as third-person knowledge but incomplete as first-person knowledge.

Back to Jackson's fable, if Mary has studied the physics of vision but has never experienced red, then she does not know everything there is to know physically about vision. Hence (b) is false and there is no contradiction. Yet, (b') can in principle be true and (d) follows. But this is not due to a failure of physicalism, namely a failure of the idea that anything that obtains is a physical fact, because (d) is not in contradiction with (a). Mary has not yet learned everything that there is to know: she has not yet gone into the physical R-state that embodies her *reflexive* knowledge of what vision is. But this is not a statement about a nonphysical reality: it is a statement about a physical reality.

Let us illustrate the situation with small example. Consider a certain number of small robots equipped with sensors, memories and a light bulb on top of each of them, moving within a finite room. In the room there is a special spot —call it the red spot— such that when a robot gets to it, its light bulb turns on and then stays on. The sensors of each robot track the position of all robots (including itself), and which light bulbs are on (including its own). Each robot has a memory where the current position of all the robots (including itself) and the state of the light bulbs, is stored. Now let's introduce the following terminology:

— The relevant *physical facts* at each time are the position of each robot, the state of the light bulbs and the state of the memories.

— A robot "*knows*" a physical fact if we can learn this fact by reading its memory.

— A robot "*knows what is like* to be at the special spot" if it has been in the special spot. In this case its light bulb is on.

Say one robot is called Mary. Then the following facts follow from the definitions: At each time, Mary knows all relevant physical facts. The fact that she has this knowledge is itself a physical fact. She can have this complete knowledge of physics facts with or without having her light bulb on. In the first case we say: "she does not know what it is like having been on the red spot". In the second case we say she does. When we say so, we are referring to a physical fact. Hence -in the sense stated- Mary can know all relevant physical facts and yet learn something new, because we are referring to different kinds of knowledge.

Of course, if we instead insist on believing that "to know what something is like" is something *over and above* being in a physical state, then the argument given does not work. But if we insist on saying so, we are not *proving* that physicalism is wrong: we are *assuming* it is. We are just saying it is.

The source of the confusion was the violation of physicalism implied by considering "knowledge" as something that cannot itself be described in physical terms. The solution of the puzzle is recognizing the role of reflexive knowledge.