Does Jackson's Knowledge Argument Succeed in Disproving Physicalism?

Introduction

In 'Epiphenomenal Qualia', Frank Jackson makes an argument against physicalism – the philosophical view that everything is physical – or, as Jackson puts it, the view that all facts are physical facts[1]. The argument begins with a thought experiment – 'Mary and the black-and-white room' – and concludes that one can come to know a new fact even when they know all physical facts about a certain field, thus intending to prove that not all facts are physical facts and that physicalism is false. This argument is known as Jackson's 'Knowledge Argument'. In this essay, I will argue that although Jackson successfully shows that Mary learns a new fact upon leaving the room, the argument is insufficient to disprove physicalism. This is because the fact which Mary learns after leaving the monochrome room can be plausibly viewed as a physical fact that Mary can learn only if she is in a p-conscious (phenomenally conscious) mental state of seeing red, making it knowable exclusively through adoption of a corresponding perspective which was unavailable to Mary prior to leaving the black-and-white room. The essay will contain three main sections: the first section features the exposition of Jackson's Knowledge Argument, the second section analyses the most prominent counter-arguments to Jackson's claim that Mary learns a new fact upon leaving the black-and-white room, and the third section presents the 'Unsatisfiable Stipulation' objection and determines whether proving that Mary learns a new fact is sufficient to disprove physicalism.

[1] Frank Jackson, "Epiphenomenal Qualia," The Philosophical Quarterly 32, no. 127 (1982): p. 127

Section I: Exposition of the Knowledge Argument

Consider the 'Mary and the black-and-white room' thought experiment. Mary is a brilliant scientist who has been in a black-and-white room her entire life. Her research leads her to know every physical fact about colours and colour-vision, such as how light interacts with the retina and what signal the brain gets from the eye when light hits it. After obtaining all the physical facts about colours and colour-vision, Mary leaves the black-and-white room[2]. Outside the room, she sees colour in real objects, such as vivid red poppies in a field.

Jackson claims that by seeing real colours, Mary learns a new fact about colour-vision, namely what it is like to see red. The 'what-it-is-likeness' of seeing red is a fact about the conscious experience of colour-vision that, in Jackson's view, cannot be obtained through physical information because nothing about the function, structure or other objective characteristic of the process of colour-vision explains what it is actually like to see red[3]. Jackson's Knowledge Argument can therefore be decomposed into the following set of premises and conclusions:

1) Physicalism claims all facts are physical facts
2) Mary knows all the physical facts about everything associated with seeing red prior to leaving the black-and-white room
3) Upon seeing the field of red poppies, Mary comes to know what it is like to see red
4) The 'what-it-is-like' quality of seeing red is a fact about colour-vision (about the experience of seeing red in particular)

---

[2] Jackson, "Epiphenomenal Qualia," p. 130

[3] Jackson, "Epiphenomenal Qualia," p. 127–31

5) Conclusion 1: Mary learns a new fact about seeing red upon leaving the Black-and-white room (From Premise 3 and Premise 4)
6) Conclusion 2: Not all facts are physical facts (From Premise 2 and Conclusion 1)
7) Conclusion 3: Physicalism is false (From Premise 1 and Conclusion 2)

Section II: Does Mary really learn a new fact upon leaving the black-and-white room?

Seeing that Premise 1 is a definition of physicalism, Premise 2 is the basic assumption that the thought experiment asks the reader to make, and Premise 3 is intuitively true, the obvious starting point for a physicalist to begin refuting the Knowledge Argument is to focus on its fourth premise. In other words, to reject the claim that Mary learns a new fact about colour-vision.

The 'Ability Hypothesis' and 'The Acquaintance Hypothesis'

David Lewis's 'Ability Hypothesis' states that Mary does not gain any new propositional knowledge (knowing-that) but rather gains new abilities (knowing-how), such as the ability to imagine and recognise red objects[4]. That way, the change in Mary's conception of the colour red can be interpreted as the acquisition of new abilities rather than as the acquisition of new factual information. If the knowledge of what it is like to see red is not propositional, then Premise 4 of the argument is false.

The Acquaintance Hypothesis, proposed as a response to Jackson by Earl Conee and later defended by Michael Tye[5], provides another counter-argument to Premise 4. The hypothesis also recognises that Mary learns something new upon leaving the room, but rejects the claim

---

[4] David Lewis, "What Experience Teaches," in Papers in Metaphysics and Epistemology (Cambridge: Cambridge University Press, 1999), p. 287–90

[5] Michael Tye, Consciousness Revisited: Materialism without Phenomenal Concepts (Cambridge, MA: MIT Press, 2009), p. 131–32

that she gains new propositional knowledge[6]. Instead, it suggests that Mary gains acquaintance knowledge, meaning that she gains direct experiential access to the colour red[7]. Though it might be sometimes possible to obtain both kinds of knowledge simultaneously, acquaintance knowledge is not factual (propositional) knowledge because it does not constitute access to any new truth about its object. Instead, it constitutes a new perspective on the same factual information that Mary already knew. Thus, the Acquaintance response likewise attempts to refute Premise 4 of the Knowledge Argument by claiming that the knowledge gained by Mary is of non-propositional nature.

Where the Ability and Acquaintance Hypotheses fail

Both of these counter-arguments fundamentally rely on certain knowledge (namely knowledge-how and knowledge by acquaintance) being categorically distinct from propositional knowledge, thus allowing new knowledge to mean something other than knowing a new fact. This categorisation, however, is where these arguments falter. I will demonstrate the inadequacy of the responses from the ability and acquaintance hypotheses in two steps. I will firstly explain why they do not show the implausibility of potential counter-arguments, which can claim that propositional knowledge has nevertheless been gained by Mary. Secondly, I will argue that the nature of the physicalist doctrine itself does not lend itself to rejecting those counter-arguments in favour of the claim that Mary only gains non-propositional knowledge upon leaving the room.

Firstly, it is not clear that knowledge-how is not just a subtype of propositional knowledge. To know **how** to swim is to know **that** to swim is like this (by moving your limbs a certain

[6] Tye expressly states that Mary becoming acquainted with seeing red provides her with "genuinely new knowledge" of the "phenomenal character to the experience of red". The importance of this new knowledge not being a physical fact, and the plausibility of this claim, is essential to the discussion in section II of this essay.

[7] Earl Conee, "Phenomenal Knowledge," Australasian Journal of Philosophy 72, no. 2 (1994): p. 144

way, positioning your torso a certain way, and so on). To know **what it is like** to see red is to know **that** to see red is like this. A similar approach can be taken to knowledge by acquaintance. To be **acquainted** with the colour red is to know **that** to see the colour red is like this. Categorisation into knowledge through ability and knowledge by acquaintance may therefore be merely a distinction in means of acquisition of that knowledge and not proof that the knowledge itself is irreducible to factual knowledge. Furthermore, even if the distinction between the proposed types of knowledge goes beyond the mere means of its acquisition, it is entirely plausible that both knowledge through ability and knowledge by acquaintance come inextricably alongside propositional knowledge. That is, gaining the knowledge that **that** is what it is like to see red might not be all there is to being acquainted with the colour red or to learning how to see red, but it nevertheless seems to, at the very least, be part and parcel of the experience of seeing red. Therefore, even if knowledge-how and knowledge by acquaintance are irreducible to factual knowledge, the ability and acquaintance hypotheses fail to prove that such propositional knowledge does not come inseparably alongside factual knowledge, which would be an equally significant issue for the physicalist.

I further argue that physicalism lacks the conceptual resources to explain how new knowledge can come without knowing new facts. At the heart of physicalism is the view that everything is physical and that all facts are made true by nothing over and above physical states. For the sake of maintaining focus, this essay will not argue for adopting a specific definition of knowledge; however, it is the overwhelmingly accepted view in contemporary epistemology that a necessary prerequisite for knowledge is truth, for what is false cannot be known. Truth, I should suppose, presumes correspondence to something actually in existence (non-propositional knowledge is still ‘about’ something that exists). Thus, knowledge by acquaintance and knowledge through ability must, according to a physicalist, have as its object an entirely physical phenomenon. In our case that phenomenon is seeing the colour

red. If knowledge by acquaintance and knowledge through ability are kinds of knowledge of seeing red, then their object is some real feature X of the physical process of seeing red (for a physicalist that is all there is). If the object of knowledge is the feature X of an entirely physical process, then to gain that knowledge is to gain the knowledge **that** (propositional) this physical process involves X. Thus, to be **acquainted** with the colour red or to gain the **know-how** of seeing red must be either identical to, or at least come inextricably alongside, knowledge **that** to see red is like X or knowledge **that** what it is like to see red involves X, which means that knowledge by acquaintance and knowledge through ability are either reducible to propositional knowledge or come inseparably alongside propositional knowledge. Therefore, Premise 4 remains unrefuted, meaning that Mary does gain knowledge of a new fact upon leaving the black-and-white room.

A possible defence from the proponents of one of the two hypotheses would be to point out that knowledge-how or knowledge by acquaintance having to correspond to a physical fact is not an issue if that correspondence is to a physical fact already known to Mary before leaving the room. They would therefore concede that non-propositional knowledge must be about a physical fact, but claim that the physical fact was known to Mary before actually seeing the colour red. What was not known to Mary, and what the real experience gave her access to, were the non-propositional aspects of the physical knowledge of seeing red that she already possessed. If this response is successful, then it would refute Premise 4 of Jackson's argument whilst avoiding the objection that for a physicalist any new knowledge must be either propositional or come alongside propositional knowledge in virtue of the need to correspond to a real physical phenomenon.

This defence, however, relocates the difficulty rather than resolving it. Whatever its form, knowledge of seeing red must answer to some real feature of seeing red. The defence must therefore specify which feature Mary's new knowledge answers to, and only two options are

available. The first is that it answers to no feature beyond those Mary already knew. It is then unclear in what sense Mary has gained knowledge of seeing red at all. What has changed is only that she now undergoes, and can recall and recognise, an experience every feature of which she had already learned about. On this option, the hypotheses successfully identify a change in Mary's relation to seeing red, not in what she knows about it, thus failing to explain Mary's learning outlined in Premise 3 in a way that is satisfactory for a physicalist. The second option is that Mary's new knowledge answers to some feature of seeing red that her knowledge before release did not capture. Since for the physicalist every real feature of seeing red is physical, Mary now knows that seeing red has a certain physical feature which she did not know it had, and this is knowledge of a new physical fact. Appealing to the "non-propositional aspects" of Mary's knowledge does not avoid this choice. An aspect of what Mary knows is either a feature of seeing red, which falls under the second option, or a feature of Mary's relation to seeing red, which falls under the first. Either way the defence fails: on the first option Mary learns nothing, and on the second she learns a new fact. The second option does, however, uncover potential for an alternative defence of physicalism which would appeal to a physical fact that Mary could not have known in the room. Section III is dedicated to analysing this defence.

Section III: Is Mary’s learning of a new fact upon leaving the black-and-white room sufficient to disprove physicalism?

The previous section of the essay established that attempts to refute Premise 4, which states that Mary’s acquired knowledge of what it is like to see red is a fact about colour-vision, have been largely unsuccessful. This is because our strong intuition that Mary gains some new knowledge upon leaving the room is difficult to explain if Mary actually learns nothing at all, and the attempts to argue that what Mary learns is not new factual information fall short of reconciling knowledge with a metaphysical reality that is compatible with physicalist

doctrine. Nevertheless, the ultimate conclusion of the Knowledge Argument – that physicalism is false – follows only if it can also be shown that the new factual information that Mary learns upon seeing the field of red poppies is not a physical fact. So far, this issue has been addressed by the second premise of the argument. Therefore, to ascertain the opposite would require refuting that premise, or at least showing that it need not be necessarily true.

'Unsatisfiable Stipulation' objection

To challenge the assumption of Premise 2, I begin by defining consciousness, in particular the two types of consciousness as described by Ned Block – access-consciousness (a-consciousness) and phenomenal-consciousness (p-consciousness)[8]. A state is a-conscious if its mental content is ready for use in logical reasoning, action or speech[9]. For example, Mary can know the properties of red poppies, such as their texture, shape, colour and biological composition, if she is in an a-conscious mental state whose content is those properties (e.g. if Mary conceptualises a red poppy before actually seeing it). A state is p-conscious when it is experiential, meaning that there is something it is like to be in that state[10]. For example, Mary grasps what it is like to see red poppies only when she is actually in the p-conscious state of seeing them in a field. This suggests that facts about p-consciousness are knowable only through the adoption of an associated perspective. This is contrasted with facts about the structure or function of things, which are objective and inferentially knowable and are not reliant on a particular point of view.

---

[8] Ned Block, "On a Confusion about a Function of Consciousness," Behavioral and Brain Sciences 18, no. 2 (1995): p. 230–31

[9] Block, "On a Confusion," p. 231

[10] Block, "On a Confusion," p. 228-230

I argue that this distinction is essential because the difference in nature between facts learned through p-conscious mental states and facts learned through a-conscious mental states is what drives our intuition that Mary learns a new non-physical fact about colour-vision upon seeing the field of poppies, when that need not necessarily be the case. When Mary actually sees the poppies, she learns what it is like to see red, or, to put it in terms more clearly showing the propositional nature of her discovery, Mary learns that **that** is what it is like to see red. Being in a p-conscious mental state of seeing red allows Mary to learn this new fact. However, that does not mean that this fact is not in itself physical. Because facts obtained through p-consciousness are only knowable through adoption of a particular perspective, the 'what-it-is-likeness' of seeing red may be a physical fact that was simply not knowable to Mary when she was confined to the black-and-white room. All the factual information about seeing red that Mary could learn prior to leaving the room was information that she obtained by occupying a-conscious mental states whose content was the description of the state of seeing red (i.e. its neural realiser, its functional role, its cause, and so on), meaning information that one can be conscious of irrespective of their specific experience and perspective. Thus, if the 'what-it-is-likeness' of seeing red is seen as a fact knowable only from the perspective of a p-conscious mental state and yet also as a physical fact about colour-vision, then it is impossible for Mary to know all the physical facts about colour-vision prior to leaving the black-and-white room. As such, Premise 2 of the Knowledge Argument might constitute an unsatisfiable stipulation, thus failing to justify the conclusion that physicalism must be false.

Possible responses to the 'Unsatisfiable Stipulation' objection and why they still do not prove Jackson right

A proponent of Jackson's Knowledge Argument could argue against the possibility of facts obtained through qualia or p-consciousness being physical facts by appealing to the characteristics of facts we know to be physical. Physical facts, such as those about the

structure of objects and processes, are objective, independent of individual experience and inferentially knowable. To claim that some physical facts lack all these properties is therefore implausible.

A more straightforward version of this argument can be made by appealing directly to Jackson's 1986 paper 'What Mary Didn't Know', where he stipulates that Mary knew "everything in completed physics, chemistry, neurophysiology, and all there is to know about causal and relational facts consequent upon all this" before leaving the room[11] . Therefore, to suggest that Mary did not know some physical fact before leaving the room is to claim that no physical science can teach some physical fact. If physicalism were true, then to have complete physical knowledge would be to have complete knowledge simpliciter[12]. Thus, a fact not learnable from complete physical sciences is just a non-physical fact wrongly called a physical fact to avoid the contradiction between its nature and physicalist doctrine.

Neither response succeeds in disproving physicalism. The inconsistency, pointed out by the first response, between the properties that we typically attribute to physical facts and the ones suggested in the 'Unsatisfiable Stipulation' objection may not be down to the difference in the nature of those facts (physical vs non-physical) but rather due to a limitation of human comprehension. The fact **that** seeing red is 'like this' is a fact about a particular mental state M. Mary's knowledge of that fact is brought about by her being in the (p-conscious) state M. Neither the fact nor the mental state M need be non-physical, even if grasping that fact requires being in M. The necessity of being in M to grasp the knowledge of what-it-is-likeness of seeing red can be seen as a property (and a limitation) of Mary's mind, not a condition that makes the fact true.

---

[11] Frank Jackson, "What Mary Didn't Know," The Journal of Philosophy 83, no. 5 (1986): p. 291

[12] Jackson, "What Mary Didn't Know," p. 291

Though it is enough for this explanation to be a mere possibility to ensure that the Knowledge Argument does not categorically disprove physicalism, I believe the use of an analogy might show it to be not only possible but also plausible. The way Rome looks from the top of the Aventine Hill is a fact about that particular hill and landscape, made true entirely by the physical layout of the land and the light reaching a specific point on the hill. To see it for oneself, one must stand on top of the Aventine Hill and look out on the city of Rome, yet one's standing there is not what makes the view as it is. Similarly, being in M is the only means by which Mary can grasp what it is like to see red, not what makes that fact true. The 'what-it-is-likeness' of seeing red can therefore be an objective physical fact which is not inferentially knowable to Mary in virtue of the limitations of human understanding and not because of the non-physical nature of the fact itself.

A similar approach can be taken to explain the inability of physical sciences to teach Mary all the physical facts about seeing red, something that the second response presents as incompatible with physicalism. If 'physical facts' referred to in Premise 2 are taken to mean facts that can be learned through physical theory alone, then that premise may well be satisfiable, but it would fail to lead to Conclusion 3, which is that physicalism must be false. This is because a physicalist is not required to concede that all facts must be teachable through physical sciences, but rather that all facts are made true by nothing over and above physical states. Mary's inability to know **that** is what it is like to see red before seeing it can be viewed as evidence of a limitation of human cognitive capacities, not evidence that this fact is made true by anything over and above physical states. Therefore, on the physicalist's reading of 'physical fact', Premise 2 is unsatisfiable if some physical facts can only be learned by a human like Mary from the perspective of a p-conscious state. Thus, Jackson is unable to conclusively claim that physicalism must be false.

Conclusion

In this essay, I outlined Frank Jackson's Knowledge Argument, decomposed it into its logical structure, and considered objections to its key premises and the degree of success with which Jackson and his proponents are able to respond to these objections. I ultimately concluded that, despite successfully proving that Mary learns a new fact after leaving the black-and-white room, the Knowledge Argument falls short of conclusively disproving physicalism. This is because appealing to the distinction between knowledge obtainable through a-conscious states whose content describes seeing red and knowledge available only from the perspective of being in a p-conscious state allows physicalists to claim that it was impossible for Mary to know all physical facts prior to leaving the room. As a result, proving that Mary did learn a new fact about colour-vision upon seeing the field of poppies is insufficient to fulfil the burden of proof necessary to conclude that physicalism is false.

Bibliography

Block, Ned. "On a Confusion about a Function of Consciousness." Behavioral and Brain Sciences 18, no. 2 (1995): 227–47.

Conee, Earl. "Phenomenal Knowledge." Australasian Journal of Philosophy 72, no. 2 (1994): 136–50.

Jackson, Frank. "Epiphenomenal Qualia." The Philosophical Quarterly 32, no. 127 (1982): 127–36.

Jackson, Frank. "What Mary Didn't Know." The Journal of Philosophy 83, no. 5 (1986): 291–95.

Lewis, David. "What Experience Teaches." In Papers in Metaphysics and Epistemology, 262–90. Cambridge: Cambridge University Press, 1999.

Tye, Michael. Consciousness Revisited: Materialism without Phenomenal Concepts. Cambridge, MA: MIT Press, 2009.