Rejecting Jackson's Knowledge Argument with an Account of *a priori* Physicalism

Reggie Mills

I. Introduction

In 1982 Frank Jackson presented the Knowledge Argument against physicalism: "Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black and white room *via* a black and white television monitor. She . . . acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red', 'blue', and so on."[1] Jackson asks, "What will happen when Mary is released from her black and white room or is given a colour television monitor? Will she *learn* anything or not?"[2] Jackson goes on to conclude, "It seems just obvious that she will learn something about the world and our visual experience of it."[3] Thus, something was missing from Mary's supposedly all-encompassing knowledge of the physical world—viz., an understanding of qualia. So, because complete physical information was insufficient for Mary's understanding of qualia, Jackson concludes that physicalism is false.

In this essay my objective will be to show that it is plausible for Mary to be able to know what it's like to have phenomenal experience while she's in the black-and-white room. I will start with some background on physicalism and present some responses Daniel Dennett has made to the Knowledge Argument. Then, I will outline the form of an *a priori* deduction of phenomenal truths from physical facts. Finally, I will show that at least some phenomenal concepts can be *a priori* deduced from lower-level properties, independent of experience. My hope is that by the end of the essay I will have made the case for *a priori* physicalism

[1] Frank Jackson, "Epiphenomenal qualia," *Philosophical Quarterly* 32 (1982): 130.
[2] *Ibid.*, 130. Emphasis Jackson's.
[3] *Ibid.*, 130.

stronger.

II. Background on Physicalism

A definition of physicalism sufficient for our purposes is that there is nothing in the world except for what is specified by P, where P is a complete description of the world in the language of physics.[4,5,6] Since physics as we currently know it is incomplete, physics under this definition would refer not to our current knowledge but to an ideal/complete physics not radically different from our own in which all physical properties and truths are known.

Physicalism is widely (though not unanimously) agreed to be a contingent fact about our world.[7,8] In other words, it is possible for worlds to exist that aren't entirely specified by P. The main reason we are led to believe that physicalism is true is causal closure: We have no way to explain interactions between physical and nonphysical objects.[9,10] Although physicalism is contingent, a world in which physicalism is true is one in which there is a necessary metaphysical connection between P and the truths that P specifies. Of interest to us are phenomenal truths (Q = all phenomenal truths). To say that P does not necessarily give rise to the truths in Q would mean that something besides P determines Q, which is inconsistent with physicalism. Further, to say that Q can be nonphysical is also inconsistent. That qualia are nonphysical is what Jackson has concluded with the Knowledge Argument (KA).

The intuition that Mary learns something upon her experience of colour is convincing, I think, because there is an ineffable aspect to phenomenal knowledge; most people with properly functioning colour-

---

[4] Frank Jackson, *From Metaphysics to Ethics: A Defence of Conceptual Analysis.* (Oxford: Oxford University Press, 1998).
[5] David J. Chalmers, and Frank Jackson, "Conceptual analysis and reductive explanation." *The Philosophical Review* 110, no. 3 (2001).
[6] Robert Kirk, *The Conceptual Link from Physical to Mental.* (Oxford: Oxford University Press, 2013)
[7] Frank Jackson, "The Case for *a priori* Physicalism," in *Philosophy-Science-Scientific Philosophy*, *Main Lectures and Colloquia of Gap 5, Fifth International Confress of the Society for Analytical Philosophy*, edited by Christian Nimtz and Angsar Beckermann (2005).
[8] Robert Kirk, *The Conceptual Link from Physical to Mental.*
[9] W.V.O. Quine, *Theories and Things.* (Cambridge: Harvard University Press, 1981).
[10] Robert Kirk, *The Conceptual Link from Physical to Mental.*

vision grasp what it's like (w.i.l.) to have a phenomenal experience of red (red_{ph}). Yet, it seems nearly impossible to provide a reductive explanation for red_{ph}. That said, it is a stretch to conclude from the KA's intuition that physicalism is false. Mary is unable to deduce w.i.l. to experience colour from P. This is an epistemic issue concerning what is knowable; it lays no claim on the metaphysical connections that make up the world. I will now outline why this is the case.

Although Q is necessarily true given P, there are two ways that Q is knowable. The first is *a priori* physicalism, which holds that Q is knowable from P independent of any additional empirical information. The second is *a posteriori* physicalism, which holds that there are some truths in Q that cannot be deduced from P. It is not inconsistent for Q to be knowable only by experience despite the necessary P–Q connection. The classic example to show what is meant by a necessary *a posteriori* truth is the H_2O–water identity: it is not an *a priori* truth that water is H_2O in all possible worlds; however, once it's discovered that water is H_2O on a particular world, anyone with a grasp of the concepts "H_2O" and "water" will be able to realize the identity's necessity.[11] So, the two possibilities for P–Q's knowability are: (1) that all phenomenal truths Q are necessary *a priori* entailed by P; and (2) that all phenomenal truths Q are necessary *a posteriori* entailed by P.

The physical properties used need not be strictly micro-level and may be things like neural organization, as long as these properties are definable in physical terms. Since in (1) you are going from lower-level physical properties up to a grasp of w.i.l. to have phenomenal experience, I will refer to an *a priori*-type deduction as "bottom-up". And, since in (2) a grasp of w.i.l. to have a phenomenal experience is gained through having the experience (experience seeming like a higher-level phenomenon relative to physical properties), I will refer to learning knowledge *a posteriori* as "top-down". Necessary *a posteriori* physicalism would explain why phenomenal knowledge Q seems ineffable and nonreductive. If phenomenal properties in Q are knowable only through experience, there would not be a logical, epistemic connection between P and Q, despite a necessary metaphysical connection – no amount of information

[11] David Chalmers and Frank Jackson, "Conceptual analysis and reductive explanation."

from P would entail Q.

Since Mary is not able to deduce phenomenal knowledge from the physical facts, the KA is compatible with necessary *a posteriori* physicalism, in which phenomenal concepts can only be learned from a top-down approach. Since, given P, phenomenal truths Q are still necessarily true, qualia are not "left out of the physicalist story", as Jackson says.[12] Thus, physicalism is not challenged by the KA.

III. Responses to the Knowledge Argument

There have been attempts to reject the intuition that Mary would learn something when she leaves the room. In *Consciousness Explained* (1991), Daniel Dennett highlights that Mary's knowing all physical information (i.e., all of P) is a prospect so immense that it is nearly unimaginable for us; as such, it is difficult for those thinking about Mary to realize what she is capable of, and for this reason Dennett calls the KA an "intuition pump".[13] He goes on to propose what Mary's response should actually be when she sees colour, according to her knowledge of P. In his scenario, Mary's captors try to trick her by presenting her with a banana that's bright blue instead of yellow. Instantly, Mary points out the trick: She knew what physical impression a yellow banana was supposed to have on her nervous system and the thoughts that would accordingly result.[14] These thoughts, presumably, would entail for Mary an understanding of w.i.l. to see yellow. The implication then is that Mary has bottom-up deduced from P w.i.l. to see yellow without ever having phenomenally experienced yellow.

More recently, Dennett has come up with RoboMary, another Mary-like scenario to help us imagine Mary's capabilities and weaken the KA's intuitions.[15] RoboMary is a Mark 19 robot with the same complete knowledge of P that Mary has. She is largely identical to other Mark 19 robots in that her mental system is capable of processing visual information and giving her colour qualia, but she has one difference: Her

---

[12] Frank Jackson, "Epiphenomenal qualia," 131.
[13] Daniel Dennett, *Consciousness Explained*, (New York: Little, Brown, 1991), 398.
[14] Daniel Dennett, *Consciousness Explained*, 399.
[15] Daniel Dennett, *What RoboMary Knows,* ed. Torin Alter and Sven Walter. (Oxford: Oxford University Press, 2007).

visual sensors—her robot-substitute for human eyes—can detect only black and white. However, the scenario goes, RoboMary studies Mark 19s with functioning, colour-capable visual sensors and the processes that give rise to their colour qualia.[16] Using this knowledge, RoboMary is able to create a prosthesis that colourizes her black-and-white visual inputs to deliver herself colour qualia.[17] I think Dennett's intended implication is that RoboMary, in studying Mark 19 colourizing processes and then creating a likeness of such processes, *understands* the connection between lower-level physical properties and phenomenal properties.

To ensure that it is clear that RoboMary is not learning w.i.l. to see colour *a posteriori* from an experience of colour, Dennett further envisions Locked RoboMary, whose colour-experience registers—the systems in the Mark 19 brain that allow colour qualia to be experienced—are locked to greyscale. So, RoboMary cannot now experience colour qualia at all—neither through visual stimulus nor imagination. But, Dennett goes on, using some free RAM in her brain RoboMary constructs a simulated model of the Mark 19 visual processing system and uses it to calculate the mental states that would normally result from the phenomenal experience of coloured objects. Dennett refers to the nonphenomenal mental state after a phenomenal experience as a "dispositional state."[18] While this dispositional state is not itself an understanding of w.i.l. to have a certain experience, it contains all the information from such an experience that would be necessary for an understanding of w.i.l. So, comparing these colour-capable Mark 19s' dispositional states to those from her own black-and-white phenomenal experience, RoboMary makes it so that after her visual experience, she gets put into dispositional states as if she were a colour-sensing Mark 19. In other words, Locked RoboMary has calculated dispositional states containing understandings of w.i.l. to have certain colour experiences. I think we are to assume here that RoboMary again understands the P–Q connection, though this time without ever having experienced actual colour qualia.

---

[16] *Ibid.*
[17] *Ibid.*
[18] *Ibid,* 24.

RoboMary has received criticism from Torin Alter.[19] The way Alter reads RoboMary is that however RoboMary puts herself into her dispositional states, whatever goes on during the "putting" step itself does not confer an understanding of the P–Q connection.[20] RoboMary discovers the relevant colour-capable dispositional states, then "comes by her phenomenal knowledge of color experience not by *a priori* deduction from physical information but rather by putting herself in a nonphenomenal dispositional state that contains the relevant phenomenal information."[21]

I sympathize with Alter's view—here is how Dennett's RoboMary scenarios make sense for me: In the first, RoboMary copies the Mark 19 computational architecture to create a prosthesis which is able to generate colour qualia from black-and-white physical-level inputs. It does not seem that RoboMary would need to understand the processes going on in the Mark 19s to be able to copy the mental structure. So, the prosthesis does the hard work of transitioning from P to Q for RoboMary; all RoboMary is conferred by the prosthesis are its outputs—colour qualia. Thus, RoboMary gains an understanding of w.i.l. to have colour experience from colour qualia and not from P, in the same way that Mary in the KA gains an understanding of w.i.l. to experience colour after leaving the room.

In the second, Locked RoboMary builds a simulation which can take her black-and-white inputs, figure out their colourized equivalent, and confer to RoboMary the dispositional brain states that normally occur after *experiences* of these colours. The information that RoboMary is using to arrive at Q is not P, but rather information about phenomenal experience. So, in both cases, Q is arrived at by top-down means and not from a grasp of P, which is no better than Mary's arrival at Q in the KA.

I hold that to reject the KA, one must show that Mary is able to *a priori* deduce Q from P (i.e., a bottom-up deduction), rather than to arrive at Q from experience (i.e., top-down). Indeed, Jackson has referred to the specifics of such a bottom-up *a priori* deduction as "*the* hard issue that

[19] Torin Alter, "Phenomenal Knowledge without Experience," in *The Case for Qualia,* ed. Edmond Wright (2008), 247-267.
[20] *Ibid.*
[21] *Ibid.,* 253.

faces physicalists today."[22] Virtually all knowledge of Q and understanding of w.i.l. to experience colour for humans is learned top-down. But, Mary does not have direct access to colour experience while in the black-and-white room. In the remainder of this essay, I will shed some light on and argue for the plausibility of a bottom-up *a priori* P–Q deduction.

IV. How *a priori* Physicalism Would Work

To be able to *a priori* deduce the truth of the conditional P ⊃ Q, one would need (a) sufficient empirical information from P such that the information implies a phenomenal concept in Q, plus (b) an understanding of the phenomenal concept in Q.[23] For example, regarding the water–H_2O identity, "if a subject possesses the concept 'water' . . . then sufficient information about the distribution, behaviour, and appearance of clusters of H_2O molecules enables the subject to know that water is H_2O, to know where water is and is not, and so on."[24] In the same vein, if P implies phenomenal truths in Q, then, given possession of a phenomenal concept such as red_{ph} and sufficient empirical information from P, Mary plausibly could deduce the phenomenal truth that "This mug looks red." The empirical information in question would include things like the ~700-nm-wavelength photons emitted by the mug's surface, the detection of said photons by Mary's red-detecting opsins and the resulting electrical signal, and, importantly, the neurological organization in the human brain which induces the phenomenal experience of redness. To deduce the connection between P and red_{ph} Mary would need to have a full grasp of all the necessary properties in P such that her grasp of these properties is equivalent to the concept red_{ph}.

Notably, none of this *a priori* deduction from P to Q involves Mary's experience of red; Mary needs only the empirical information from P and a grasp of the concept red_{ph}. Thus, the deduction is *a priori*/bottom-up. I do not know the specifics of the P–Q deduction, so the "hard issue" still remains. But, given that the connection from P to Q is metaphysically

[22] Frank Jackson, "The Case for *a priori* Physicalism, 264.
[23] David J. Chalmers and Frank Jackson, "Conceptual Analysis and reductive explanation."
[24] *Ibid.*, 323.

necessary, P–Q entailment seems plausible.

It is key to realize that the amount of information from P that Mary will need in order to have a grasp of truths in Q will not be trivial. Of particular note would be the neurological organization that gives rise to phenomenal experience and to the possession of phenomenal concepts. Whatever goes on here is something we evidently do not currently understand. Part of the reason for this, it seems, is that we just do not have enough empirical information about the brain to put together a cohesive explanation of consciousness; what we know from P is not enough to imply Q. For example, for H_2O–water, it would be difficult to deduce water from properties of H_2O molecules if we did not know about the fundamental forces involved. But, it is also possible that no human will ever be able to have a grasp of properties from P that would entail phenomenal truths in Q. As Dennett says regarding his blue banana example, Mary knowing w.i.l. to experience blue "wasn't easy. She deduced it, actually, in a 4,765-step proof."[25] However, human incomprehensibility does not limit the apriority of the P–Q entailment. "Apriority concerns what is knowable *in principle*."[26] Consider also the apriority of H_2O–water: Even with an understanding of the concept "water", any human would be hard-pressed to deduce truths relating to water from empirical information about H_2O molecules. But, the H_2O–water identity is still entailed *a priori*, and is routinely calculated via computer simulations.

The *a priori* deduction of Q from P that I showed above was presented with Mary already possessing a phenomenal concept such as red_{ph}, but the question we are interested in is if Mary could deduce w.i.l. to experience red from P without any pre-existing grasp of red_{ph}. So, we just need to rephrase the *a priori* deduction such that Mary uses sufficient empirical information from P to arrive at an understanding which would be equivalent to a grasp of the phenomenal concepts involved. Just as with H_2O–water, there will be a point at which a macroscopic grasp of the micro properties involved (from P) will be equivalent to a grasp of the higher-level concept (water or red_{ph}). In this way, possession of the concept red_{ph}

[25] Daniel Dennett, *What RoboMary Knows*, 16.
[26] David Chalmers and Frank Jackson, "Conceptual analysis and reductive explanation," 334. Emphasis mine.

is not a prerequisite for a bottom-up arrival at an understanding of the concept.

V. Phenomenal Knowledge without Experience

Part of the difficulty, I think, in imagining an arrival at phenomenal concepts without experience is that virtually every phenomenal concept we possess has been gained *a posteriori*. Let me here clarify that a grasp of a phenomenal concept is exclusive from the phenomenal experience itself. Think of the phenomenal concept of w.i.l. to experience pain ($pain_{ph}$). The thought of $pain_{ph}$ is not tied to an experience of pain, despite $pain_{ph}$ containing a full grasp of w.i.l. to experience pain. Similarly, understandings of other phenomenal concepts do not contain phenomenal experience. For taste and smell concepts I think the parallel to pain is obvious. For visual and auditory concepts, though, it seems that thinking of the concepts are almost unconscious triggers for imagining these concepts' phenomenal experiences. But, it is still possible with effort to think of these types of concepts without imagining or experiencing them. So, to reiterate, possessing a phenomenal concept is sufficient for understanding w.i.l. to have the relevant phenomenal experience. Then, a deduction of a phenomenal concept from lower-level properties would be sufficient to understand w.i.l. to have the corresponding phenomenal experience; nothing more would be learned by having the relevant experience once a concept is possessed.

Now we need to show how such an *a priori* (bottom-up) deduction of phenomenal concepts from lower-level properties is possible—i.e., tackle the hard issue. Here is an example to show at least that certain phenomenal concepts are bottom-up deducible: Someone familiar with a quadratic equation in the form $y = x^2$ can easily understand the parabolic shape of the function's curve, how changing the function to $y = 2x^2$ will make the parabola narrower, etc. This understanding is bottom-up deduced; we can change the equation to something novel and not previously experienced without compromising the apriority of the deduction. The phenomenal concept of the curve's shape ($curve_{ph}$), I think, is a simpler case in the many examples of phenomenal concepts. As I mentioned earlier, the *a priori* deduction of many of these phenomenal concepts (ones like red_{ph} and $pain_{ph}$) from P would be quite difficult. I at

least think that the *a priori* deduction of $curve_{ph}$ from a quadratic equation is evidence that when lower-level properties contain higher-level phenomenal properties, a grasp of the lower-level properties can be equivalent to (and hence entail) an understanding of w.i.l. to experience those higher-level phenomenal properties. It should also be noted that $curve_{ph}$ is deduced from properties in the language of mathematics, an entirely different language from that of phenomenal concepts. So, transitioning between the potentially different languages of physical properties in P and phenomenal properties in Q should not affect the apriority of a P–Q deduction. Thus, I think it sounds plausible that Mary could have a grasp of properties in P equivalent to phenomenal concepts in Q without having had the corresponding experience for any Q-concepts, and this sort of P–Q *a priori* deduction would not be lacking any of the knowledge necessary for Mary to understand w.i.l. to experience phenomenal concepts.

VI. Conclusion

As I have tried to defend, I think it is possible in principle to *a priori* deduce all phenomenal truths and concepts from P. So, when Mary leaves the room, phenomenal concepts and knowledge of w.i.l. to experience these concepts will already be familiar to her; she would gain no knowledge. I have shown that it is possible to *a priori*, bottom-up deduce the phenomenal concept of a quadratic curve from a quadratic equation. However, I have no idea how the *a priori* deduction of complex phenomenal concepts like red_{ph} and $pain_{ph}$ from P would work, and I leave the "hard issue" of consciousness unsolved for these. My belief is that the difficulty here is a result of human limitations, not a reflection of the metaphysical relationship between the physical and mental. I see there to be two possibilities for the current limitation: (1) That our understanding of the phenomenal concepts involved is insufficient, and (2) that we lack some of the necessary empirical information in P. Based on our current knowledge of the neurological organization of the brain I think (2) is undeniably true. Regarding (1), I am unsure if concepts such as red_{ph} as we understand them are sufficient for a P–Q deduction. And, (1) and (2) need not necessarily be mutually exclusive; a possession of the phenomenal concepts involved sufficient to make the *a priori* deduction

of Q from P would include a grasp of the appropriate empirical information from P.

Works Cited

Alter, Torin. "Phenomenal Knowledge without Experience." In *The Case for Qualia*, ed. Edmond Wright, (2008): 247–267.

Chalmers, David .J. and Jackson, Frank. "Conceptual analysis and reductive explanation." *The Philosophical Review*. 110, no. 3(2001):315-360.

Dennett, Daniel. *Consciousness Explained*. (New York: Little, Brown, 1991).

—— *What RoboMary Knows.* in *Phenomenal Concepts and Phenomenal Knowledge*, ed. Torin Alter and Sven Walter. (Oxford: Oxford University Press, 2007).

Jackson, Frank. "Epiphenomenal qualia." *Philosophical Quarterly*. 32(1982):127–136.

—— *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press, 1998.

—— "The Case for *a priori* Physicalism." In *Philosophy-Science-Scientific Philosophy, Main Lectures and Colloquia of Gap 5, Fifth International Congress of the Society for Analytical Philosophy*, ed. Christian Nimtz and Angsar Beckermann. Mentis. (2005).

Kirk, Robert. *The Conceptual Link from Physical to Mental*. Oxford: Oxford University Press, 2013.

Quine, Willard van Orman. *Theories and Things*. London: Harvard University Press, 1981.