Introduction: The Enduring Significance of Jackson's Knowledge Argument

Sam Coleman

I.1

In 1982 Frank Jackson published 'Epiphenomenal Qualia', in which he imagined the near-future scenario of a researcher into colour vision who is confined to a monochrome environment. Because it is the imaginary near future, she is able to compile *all* the scientific information about the physical goings-on that underpin colour vision. He describes her situation thus:

> [A] 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 specializes in the neurophysiology of vision and 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. She discovers, for example, just which wavelength combinations from the sky stimulate the retina, and exactly how this produces *via* the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence 'The sky is blue'.
>
> [B] What will happen when Mary is released from her black and white room or is given a color television monitor? Will she *learn* anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. (Jackson 1982: 130)

In a later paper Jackson envisages that on leaving the black-and-white room Mary 'will learn what it is like to see something red, say',[1] and philosophers

[1] Jackson 1986: 291.

have associated Mary with learning what red is like ever since. Given that Mary does learn, Jackson next infers:

> [C] [I]t is inescapable that her previous knowledge was incomplete. But she had *all* the physical information. *Ergo* there is more to have than that, and Physicalism is false. (Jackson 1982: 130)

In this way Mary's story is used to deploy an argument against *physicalism*, roughly the view that the world is wholly physical. The argument moves from considerations about Mary's knowledge to the conclusion of physicalism's falsity. Hence Jackson dubbed it 'the knowledge argument'. Generalising the knowledge argument's purport, he continued:

> Clearly the same style of Knowledge argument could be deployed for taste, hearing, the bodily sensations and generally speaking for the various mental states which are said to have (as it is variously put) raw feels, phenomenal features or qualia. The conclusion in each case is that the qualia are left out of the physicalist story … the polemical strength of the Knowledge argument is that it is so hard to deny the central claim that one can have all the physical information without having all the information there is to have. (Jackson 1982: 130)

We can put the knowledge argument more formally by compressing the passages [A], [B], and [C] into two premises and a conclusion:

Premise 1 Mary before her release knows everything physical there is to know about people.

Premise 2 Mary before her release does not know everything there is to know about people, because on her release she learns what it is like for them to visually experience red.

Conclusion There are things to know about people that escape the physicalist story, so physicalism is false.

This formulation is slightly adapted from Jackson (1986: 294). There Jackson's premises talk of Mary knowing everything physical about *other* people; but the conclusion is that there are truths about other people *and herself* that escape the physicalist story. It is clear enough that there is something Mary does not know about herself pre-release: she may know what physical brain state she will come to have when she sees red, but she does not know what her associated visual experience will be like. So I talk here simply of 'people', to include other people and herself. Jackson also omits mention of the sort of thing Mary will learn, in premise 2, so I have made

that explicit. Moreover, Jackson's original conclusion does not say openly that physicalism's falsity follows, though this inference is clear from the rest of his exposition – so I have again spelt that out. Finally, his conclusion talks of truths, not knowledge. But the sense of 'knowledge' intended throughout must be of truths (or facts) for the argument to work (a point objectors focus on – see below), and *knowledge about* is plausibly of this sort (more on this point below).

A great deal has been written about Jackson's knowledge argument in the nearly forty years since its publication; so much, indeed, that if one keeps up with the literature one is apt to start to feel – a bit like Mary herself – that one knows all that can possibly be written down on the subject. Still, it seems the way with philosophy that there is always more to say. New decades bring fresh ideas, or at least a freshening of approaches to existing ideas, and that recurrent, curlicuing, self-updating process is the main manner by which large-scale progress is achieved in philosophy, when it is achieved at all. Therefore it is worth while to revisit this classic argument now – in order to witness how the years have affected it, to understand how the argument has affected us in those years, and to anticipate where it might lead us henceforth. Herein is a new set of essays on the knowledge argument, embodying major trends of thought at the present time. This volume is a record of the argument's relevance to and impact upon the present, and a pointer to its enduring significance in the future.

The knowledge argument shares those virtues characteristic of the very greatest arguments in the history of philosophy, possessing an overt formal simplicity of presentation, and invoking intuitively accessible ideas, but combining this with a depth of reach into complex fundamental issues. One thinks also of Descartes's *Cogito* and Anselm's ontological argument in this bracket, to name but two. Jackson's argument is additionally notable for being doubly influential. A horde of philosophers have been impelled to react to its *content*, either by defending or debunking its reasoning and the claims Jackson took to follow from this. (Interestingly, Jackson himself is now a debunker, having undergone a doctrinal turnaround every bit as dramatic, philosophically speaking, as Asoka's conversion to Buddhism. Nowadays a physicalist, his chapter in this collection presents his latest thoughts on where earlier Jackson went wrong.) But philosophers have also been influenced by the knowledge argument's *form*, proceeding to adapt the scenario of an apparently all-knowing subject who nonetheless learns in order to demonstrate conclusions not necessarily directly related to claims about consciousness and physicalism. Essays of the two kinds feature in this collection. In both ways,

the knowledge argument's circle of influence is far from completed. It remains as timelessly timely as ever: a mirror of perennial concerns about mind, consciousness, and matter, as well as a crucible for the forming and testing of new philosophies – still a key resource in our endeavour to understand our place as minded creatures in a world of matter.

I.2

The ostensible target of Jackson's argument is *physicalism*, roughly the thesis that every concrete existent in our world is wholly physical in nature – i.e. all properties are physical properties, and these are the only properties possessed by the things that bear properties. Mary's situation seems to show that one could know all about the world's physical nature and yet learn still more about it, specifically as regards the visual colour experiences people routinely undergo.

The physicalist thesis can be put somewhat more precisely by saying that a duplicate of our world that featured only its physical ingredients would be its duplicate in every way – with all the aesthetic, moral, meteorological, and, most notably, psychological richness of actuality.[2] What does 'physical' mean? Without facing the myriad complications of pinning down this term,[3] the ruling idea in the literature on consciousness and the knowledge argument is that the physicalist looks to science, especially the hardest science, physics, for the catalogue of basic physical existents and a description of their characters and the events they participate in. The controversial physicalist claim when it comes to the mind, then, is that it ultimately involves nothing but physical goings-on in this sense.

The knowledge argument is important in part because physicalism is such a large target. Though it has never totally dominated philosophical opinion, it certainly represents the orthodoxy in the philosophy of mind over the last century or so (the thesis that the mind is material of course has a long history).[4] How did physicalism become prevalent? A prominent motivation, which acquired its full strength with the physical, biological, and neuroscientific advances of the last century,[5] derives from the 'causal

[2] See Jackson (1993) for the first formulation of this idea, but further complications are discussed in Stoljar (2015a) and references contained therein. See also Montero (2012).
[3] See e.g. Crane and Mellor (1990); Spurrett and Papineau (1999); Wilson (2006).
[4] For twentieth- and twenty-first-century physicalism see Smart (1959); Armstrong (1968); Lewis (1996); Melnyk (2003). For historical proponents see e.g. Democritus; Hobbes (1655).
[5] See Papineau (2002) for an account of this development in connection with the causal argument.

argument' for physicalism, centred on the empirical premises that physical events always have purely physical causal histories, and that conscious mental states routinely have physical effects, e.g. bodily motions. From this it follows, with a little more reasoning,[6] that the relevant conscious mental states are themselves physical. There are many ways of resisting this argument,[7] but it has doubtless been, as much implicitly as explicitly, an influential driver to physicalism.

As stated, physicalism is an *ontological* creed: about the nature of what there is. Yet Jackson's case against it turns on considerations concerning *knowledge*.[8] These facts are especially evident in the conclusion as I framed it, which contains an inference – the first part summarising the knowledge claims made in the premises, and the second part drawing the metaphysical moral of physicalism's falsity. Philosophers of the present era are deeply wary of deriving metaphysical consequences from epistemic claims.[9] In order to generate his metaphysical conclusion, Jackson must therefore attribute to physicalism an epistemological thesis as a corollary to its metaphysical claim. This thesis is evidently that the physical truths are the only truths one need know – science can tell us all the facts about the world, in other words. More strictly put, the content of any apparently non-physical truths (truths in non-scientific vocabulary) can be derived *a priori* given full physical knowledge and relevant empirical premises. If physicalism is committed to this thesis, and yet Mary can learn more about the world on seeing red, despite knowing all the scientific facts, then physicalism's falsity as a metaphysical doctrine plausibly

[6] A 'no systematic overdetermination' premise is also usually inserted, to rule out the possibility that a non-physical mental and a physical event are always sufficient but independent causes of the physical effects of conscious mental states.

[7] Popular among these are arguments that the physical completeness premise is as a matter of fact false, or begs the question when formulated precisely (see Gibb (2014) for a survey). 'Russellian monism' is another way out, for those who consider it a form of physicalism – see e.g. Chalmers (2015); Montero (2015); Goff (2017) and this volume.

[8] The 'knowledge intuition', that someone can know all physical or scientific truths and yet learn about experience is, as has often been noted, of great vintage, and Jackson's version has antecedents in Broad (1925); Dunne (1927); Russell (1927a); Nagel (1974); Robinson (1982), and others. See the excellent introduction to Ludlow, Nagasawa, and Stoljar (2004) for further sources of historical arguments of the same style, as well as discussion of the ways that Jackson's argument improves upon them, and Strawson (this volume) for erudite and entertaining exposition of several more. Jackson's influential revisiting of the knowledge intuition in 1982 is an example of philosophy's self-updating process, as mentioned in I.1 above.

[9] Kripke (1972) was influential in this separation of matters epistemic from matters metaphysical.

follows. Jackson has expended much energy, especially since his conversion to physicalism, in defending this epistemological corollary to physicalism.[10]

The knowledge argument has a deceptively smooth philosophical hide, but this appearance conceals a complicated set of interlocking theoretical vertebrae, and physicalists (and others) have been quick to attack each joint in the attempt to break the argument's back. Some points of attack: (1) Need Mary learn at all, given the (for us) unimaginable richness of her physical knowledge? (the *no-new-knowledge response* – also known as the *ignorance response*[11]). (2) Indeed, might the intuition that Mary learns be based upon subtle misconceptions about the nature of colour experience? (*the representationalist response*[12]). (3) Assuming that she does learn, is it so obvious that she learns a *truth* or *fact* about the world, or might she just gain a set of *skills* or *abilities* which come only with experiences, e.g. to remember and recognise the novel visual experience of red? (*the ability hypothesis*[13]). In that case, the physicalist need not concede that the scientific account of the world is incomplete. (4) And even if she learns a truth or fact, must it concern a new subject matter – some non-physical property of consciousness, as Jackson holds – rather than being in some sense a mere re-phrasing or re-conceptualising of knowledge already in her possession under a scientific guise? (*conceptual dualism*, aka *the phenomenal concept strategy*[14]). In that case physicalism is not guilty of a lack of metaphysical coverage, even if alternative vocabularies also exist for expressing some of the physical facts. (5) Or, then again, might Mary's gain in knowledge consist merely in a hitherto-unavailable, and peculiarly *direct*, sort of cognitive grasp of visual redness, a grasp comporting by itself no new factual content? (*the acquaintance response*[15]). (6) For that matter, is it even so clear that physicalists must hold that all physical truths can be known via the relatively remote means at Mary's disposal within her room? (*subjective*

[10] See e.g. Jackson (1998a), (2005a). His present position, roughly, is that Mary can work out what it is like to see red ahead of time, on pain of physicalism's falsity – see his contribution to this volume.

[11] For example Churchland (1985); Dennett (1991); Stoljar (2006). See McClelland's contribution to this volume.

[12] Jackson (2003), this volume.

[13] See e.g. Nemirow (1980); Lewis (1990); Mellor (1993); Jackson (2003). Objections: Jackson (1986); Stanley and Williamson (2001); Coleman (2009b). See Kind's chapter in this volume.

[14] Loar (1997); Papineau (2002); Levin (2007a); Balog (2012). Objections: Chalmers (2006); Levine (2007); Goff (2017). Defence: Diaz-Leon (2010). Anti-physicalist versions: Gertler (2001); Chalmers (2003). See Goff's chapter in this volume.

[15] Conee (1994); Tye (2009). Gertler objects (this volume) that acquaintance is unhelpful to physicalists. See also Goff 2015a; Pitt (this volume).

physicalism[16]). These questions, and combinations thereof, are among the major issues in the vast debate around the knowledge argument, and they all surface in one or another way in the following chapters.

I.3

Tim Crane (Chapter 1, 'The Knowledge Argument Is an Argument about Knowledge') argues that the knowledge argument has an exclusively *epistemic* payoff – in fact nothing metaphysical follows from it. Therefore, it does not threaten physicalism after all. What it shows, Crane maintains, is only that there is some factual knowledge that one cannot have without undergoing specific kinds of experience. Criticising prominent physicalist objections to Jackson's argument, Crane makes the case that it is nonetheless an ineffective tool in the hands of anti-physicalists. One benefit of Crane's exposition is his bringing to the surface a premise often left tacit in presentations of the knowledge argument, including Jackson's own: that Mary learns a fact. Still, the notion of fact is ambiguous between a metaphysical and a purely epistemic sense, and Crane gives reasons for favouring an epistemic reading. But this reading does not threaten physicalism, for, according to Crane, physicalists should not hold that all pieces of knowledge are knowable through science. Crane's chapter heads the collection not least due to its clear and full account both of the knowledge argument and of some major physicalist objections, features that make it ideal for orienting, or re-orienting, oneself in the debate.

David Rosenthal (Chapter 2, 'There's Nothing about Mary') reveals a tension among the claims that Mary's knowledge is new, and that it is factual. Through a critique of existing treatments of the knowledge argument, notably the phenomenal concepts response,[17] the ability hypothesis,[18] and the acquaintance response,[19] Rosenthal argues that Mary's knowledge, if factual, must be available to her within her black-and-white room. Alternatively, if she encounters a genuine epistemic novelty, this cannot involve factual knowledge (only something like acquaintance). Either way the knowledge argument's conclusion,

[16] See e.g. Searle (1992); Crane (2003, this volume); Van Gulick (2004); Howell (2009a); Goff (2017). I include here 'Russellian monism' as a response to anti-physicalist arguments including the knowledge argument. See Goff's chapter in this volume for Russellian monism, and the summary of it below, as well as Alter and Nagasawa (2012) and Chalmers (2015). Russellian monism need not be construed as physicalism, but proponents often do so construe it, and it is certainly physicalism in the sense of subjective physicalism.

[17] See n. 14 for references.

[18] See n. 13 for references.

[19] See n. 15 for references.

that Mary learns a new truth about people, fails to be secured. Rosenthal further criticises a consciousness-based way of construing the subjective content of conscious experiences, and offers an alternative based on his 'quality-space theory'. This approach has interesting consequences for Mary's first experience of red: it may be far less rich and far less like any conscious experience of ours than commentators have assumed.

Brie Gertler (Chapter 3, 'Acquaintance, Parsimony, and Epiphenomenalism') explores the implications of the acquaintance response, the suggestion that Mary cannot know what red is like without experiencing it because knowledge of experiential properties requires a special kind of direct cognition – known as acquaintance.[20] This response takes physicalist as well as dualist forms, but Gertler argues that embracing acquaintance reduces physicalism's appeal with respect to dualism. That causes a problem for physicalism, she suggests, because invoking acquaintance is an attractive way of analysing what happens to Mary. Not only is acquaintance problematic for physicalism in itself, Gertler argues, but positing it also bolsters the epiphenomenalist variety of dualism, on which experiential properties are causally inert with respect to the physical.[21] Since acquaintance does not present experiential properties *as* physical – for otherwise science would tell Mary what red is like – Gertler finds acquaintance physicalists to be deeply pessimistic about our conceptualisations of experience: such conceptualisations, they must hold, are deceptive about the real nature of experiential properties. On the other hand, acquaintance physicalists are highly *optimistic* about scientific conceptualisations of the world. Since physicalism's claim to greater parsimony than dualism rests on this optimistic attitude about scientific concepts coupled with pessimism about experiential concepts, Gertler observes that this apparent double standard requires justification independent of physicalism's truth. Having issued this challenge to physicalists she follows up by wielding acquaintance to deflect several important objections to epiphenomenalism.

David Pitt (Chapter 4, 'Acquaintance and Phenomenal Concepts') argues that knowing what an experience is like is pure acquaintance knowledge, not to be construed as at all propositional or conceptual. Rather, Pitt claims, knowing what an experience is like is to be identified simply with *having* that experience. It follows that when Mary learns what red is like, that knowledge involves no propositional epistemic gain. This view is similar to Conee's,[22] but

[20] For references see n. 15.

[21] This is the position originally embraced by Jackson (1982) in his work, hence the title 'Epiphenomenal Qualia'.

[22] Conee (1994).

Pitt motivates it via an original critique of the phenomenal concept strategy.[23] He argues that there is no principled way of making out the constitutive role that experiences are supposed to play in providing the contents of phenomenal concepts – hence there are no such concepts. It follows that there are no thoughts about experiences that one can think only having had the relevant experiences. As Pitt puts it, the difference between Mary and someone who has seen red is *perceptual*, not *conceptual*. Pitt maintains that his account is neutral about the nature of the property Mary encounters when she experiences red, and about experiential properties in general. But as a way of overturning physicalism, his reasoning implies that the knowledge argument fails.

Frank Jackson (Chapter 5, 'The Knowledge Argument Meets Representationalism about Colour Experience') further develops his response to the knowledge argument based on representationalism about perceptual experience.[24] He thereby rejects not only his earlier argument for dualism, but his even earlier arguments for a sense-datum theory.[25] Jackson now defends a 'possible worlds'-based account of the content of visual experience, analysing the vaunted 'feel' of visual experience as comprising the conjunction of a certain seamless richness, 'pig-headedness' (even when we know we are witnessing an illusion the illusory appearance persists), its nagging quality (the illusory appearance insists that it represents how things are), and a striking immediacy. On this account, Mary should be able to deduce the content of her experience of red given her physical knowledge. For a physicalist account of the representational content of experience should be possible, Jackson observes – and if it is not, the knowledge argument is not needed in order to refute physicalism. He concedes a sense in which Mary cannot know what red is like, but explains away this residue in terms of the ability hypothesis,[26] and ends by offering an analysis of the content of Mary's new visual state, as representing that there is a property of surfaces standing in the resemblance relationships characteristic of red as captured by the colour solid. Breaking with previous work, he defends the thesis that such properties are genuinely instantiated.

Galen Strawson (Chapter 6, 'The Mary-Go-Round') argues that though the descriptive reach of physical science is not sufficiently extensive to include what it is like to have visual colour experience, this supports a broadening in our conception of physicalism, rather than the inference that physicalism is false.

[23] See n. 14 for references.
[24] Jackson (2003).
[25] Jackson (1977). It is a superlative career indeed wherein one can successively, and influentially, repudiate various pieces of one's own finest work.
[26] See n. 13 for references.

In Strawson's view the moral of Mary's story is that physicalism should not be conflated with 'physics-alism', on which the physicalist ontology is restricted to those things and properties that physics – and the physical sciences more generally – can comprehensively characterise. He diagnoses this conflation as at fault for mistakes on both sides of the 'Mary-go-round': namely, a physicalist tendency to deny that Mary learns about the world when she experiences red, and the equally culpable propensity of anti-physicalists to take the Mary story as refuting physicalism. Mary, thus, is no problem for physicalism, properly understood. The quasi-Kantian lesson Strawson draws is humility about the nature of the physical, excepting that part of the physical present in our experiences.

Though it was once a popular thesis that Mary's learning consists at least in part in gaining a new 'phenomenal concept' – a concept directly picking out the experiential character of visual redness,[27] a widespread view nowadays is that the relevant concept is not beyond Mary's reach within her black-and-white room. For Mary can acquire the term RED, or even PHENOMENAL RED, from her books, or community, and use it to refer to what other people refer to with it. This constitutes an objection both to the physicalist phenomenal concept strategy, and to the knowledge argument itself, on the assumption that the argument implies that Mary gains a new concept. In response, the move has been to say that although Mary may possess this concept, she cannot *fully* possess, or master, it without experiencing red. Torin Alter (Chapter 7, 'Concept Mastery, Social Externalism, and Mary's New Knowledge') defends the thesis that Mary's epistemic progress consists at least partly in gaining mastery of the phenomenal concept RED against objections by Ball and Rabin. What those objections show, he argues, is that Mary's original story might have to be modified if purveyors of phenomenal concepts are to establish the existence of a gap between the physical and conscious experience. Specifically, proponents of the knowledge argument might have to consider a Mary who is in full possession of the phenomenal concept of red. But such methodological concessions do nothing to blunt the force of the knowledge argument, so Alter maintains.

Amy Kind (Chapter 8, 'Mary's Powers of Imagination') takes issue with the ability hypothesis – the proposal that rather than learning factually about people, Mary's new knowledge of what red is like is best analysed as a gain of abilities: to recognise, remember, and imagine red.[28] Kind

[27] See n. 14 for references.
[28] For references see n. 13.