Chapter 3 – The Knowledge Argument

In the first half of this book I will press a familiar charge: Physicalism is false because it cannot account for consciousness. If physicalism were true there would be no consciousness. There is consciousness, therefore physicalism must be false.

There are two much discussed ways of pressing this charge: the knowledge argument and the conceivability argument. In this short chapter I discuss the first of these arguments. The knowledge argument, as the name suggests, focuses on the relationship between knowledge of physical truths on the one hand and knowledge of truths about conscious experience on the other. It provides a strong case that there is an epistemic gap between these two kinds of truth, in the specific sense that one could not deduce any experiential truths about the world from knowledge of the physical truths about the world.

However, physicalism is a metaphysical, not an epistemological thesis. It is the thesis that all of reality – including experiential reality – is constitutively grounded in physical reality. It is not the thesis that all knowledge – including knowledge of experiential truths – can be derived from knowledge of physical truths. Consequently, although ultimately I think that physicalism can be refuted, I don’t think that the knowledge argument, in and of itself, has the resources to do this.

Superficially, my target in chapter 3-5 is not physicalism as such, but what I called in the last chapter pure physicalism: physicalism in conjunction with the thesis that fundamental reality can be exhaustively described in the mathematico-causal vocabulary of physics. I do this simply because the arguments of this chapter are more easily stated with pure physicalism rather than physicalism as the target. When in chapter 6 we have got a clear definition of ‘proto-mentality’, and of the distinction between physicalism and Russellian monism, I will show how the case against pure physicalism built in the first half of the book is also a case against physicalism in general.
3.1 Black and white Mary

The knowledge argument against physicalism goes back at least to Leibniz, but its most well-known contemporary formulation is due to Frank Jackson. Jackson’s argument is expressed with a story concerning a woman called Mary, who has gone on to become one of the most discussed non-existent characters in philosophy. The story of Mary goes as follows.

Mary is a brilliant neuroscientist who specialises in colour vision. In fact, she knows everything the physical sciences can tell us about what’s goes on in human brains when people see colours. For example, she knows what physical processes arise from the impact of different wavelengths of light on the retina, and how these physical processes give rise to various forms of behaviour, including verbal reports such as, ‘Ah, what a lovely shade of red.’ However, for some reason that’s never made clear, she’s spent her entire life in a black and white room. Despite knowing so much about the science of colour vision, she’s never actually experienced any colours other than black and white and shades of grey.

At the climax of our story, Mary is liberated from her black and white prison and for the first time sees something red, perhaps a bright red rose lying on the ground outside. According to the proponent of the knowledge argument, when Mary sees red for the first time she learns something new, namely *what it’s like to see red*. Despite her extensive neurophysiological knowledge, Mary’s encounter with red gives her a new kind of knowledge. The fact that Mary can learn something new about colour experience despite knowing all that the physical sciences can teach about it, is supposed to show that there is more going on in colour experience than the physical sciences are able to describe. Pure physicalism must, therefore, be false.

____________________

1 Leibniz 1714: section 17; Jackson 1982, 1986. In the same year that Jackson published his argument, Howard Robinson (1982) published a very similar version of the knowledge argument, but involving sounds rather than colours.
3.2 Responses to the knowledge argument

There are broadly speaking three kinds of response to the knowledge argument, each of which concedes a little more to the argument, whilst not quite conceding the falsity of physicalism:

- **The no-compromise response** – Mary learns nothing new when she leaves her black and white room.
- **Non-propositional knowledge responses** – Mary learns something new when she leaves her black and white room, but what she learns is not propositional knowledge, but *know how* or *knowledge by acquaintance*.
- **The new truth/old property response** – Mary learns new propositional truths when she leaves her black and white room, but those truths are simply new ways of thinking about properties and states of affairs she already knew about whilst in her room.

In what follows I will argue that:

- When fully spelt out, the no-compromise response involves some extremely implausible claims, and thus is not an adequate response to the knowledge argument.
- Whilst Mary surely does learn some new know how and some new knowledge by acquaintance when she sees red for the first time, there is strong reason to think that she also gains some new propositional knowledge. Hence, the non-propositional knowledge responses are not adequate responses to the knowledge argument.
- The new truths/old property response constitutes an adequate response to the knowledge argument, in the sense that the knowledge argument does not have the resources to render it implausible.
3.2.1 The no-compromise response

In *Consciousness Explained* – commonly referred to by critics as ‘Consciousness Denied’ – Daniel Dennett refuses to accept that there are any grounds for denying that Mary would not be able to work out, from the pure physical facts, what it’s like to see red:

It is of course true that in any realistic readily imaginable version of the story, Mary would come to learn something, but in any realistic, readily imaginable version she might know a lot, but she would not know everything physical. Simply imagining that Mary knows a lot, and leaving it at that, is not a good way to figure out the implications of her having ‘all the physical information’ any more than imagining she is filthy rich would be a good way to figure out the implications of the hypothesis that she owned everything.

Dennett’s response plays on the incomprehensible enormity of knowing *all* the physical truths, down to the tiniest details of each individual field or particle, encouraging the reader to think herself foolish for speculating about what would follow from such an unimaginable cognitive state. However, although this reflects how Jackson sets things up, it’s not necessary to hold that Mary knows *all* the physical truths, down to the level of fundamental physics.

Firstly, it is a reasonable assumption that the physical facts relevant to consciousness – at least if physicalism is true – obtain at the neurophysiological level; facts at the level of quarks and electrons will not be relevant. More generally, let us reflect for the moment on what would have to be the case for knowledge of what it’s like to see red to be deducible from pure physical truths. Pure physical truths are truths about causal structure: truths that

\[ \text{________________________} \]

---

2 Dennett obviously does not use my terminology of pure physical facts, but I think it’s pretty clear from context that this is what he’s talking about.

3 Dennett 1991: 400.
can be specified in the austere mathematico-nomic vocabulary of the physical sciences. Plausibly from such truths about causal structure one can deduce only more truths about causal structure. Thus, if the knowledge of what it’s like to see red can be deduced from causal-structural truths, then what it’s like to see red must itself be a kind of causal-structural truth. Furthermore, it must be a fairly simple truth concerning the causal role ordinary people associate with red experiences; as ordinary people, including young children, know what it’s like to see red.

It seems then that the no-compromise response must be committed to some form of analytic behaviourism or functionalism: the view that all truths about mentality, including truths about what it’s like to have experiences (assuming there are such truths) concern behaviour or behavioural functioning. Truths about pain, for example, are really truths about pain behaviour, or inner states that are disposed to instigate pain behaviour as the result of bodily damage. If knowledge of what it’s like to see red can in principle be deduced from the pure physical facts of the brain, and is knowledge which is had by children, then it must be knowledge concerning the behavioural-functional role which ordinary people associate with red experiences. At least, I can see no other way to make sense of the no-compromise response.

Appreciating this reveals that the knowledge argument need not be run with such a farfetched story as Jackson offers us. If the meaning of ‘This is what it’s like to see red’ is given by the behavioural-functional role ordinary people associate with red experiences, then Mary doesn’t need to know every last physical detail of what’s going on in the brain in order to come to know that truth. She merely needs to observe, on her black and white television, how people use language relating to red experiences, so that she can get a grip on the causal role ordinary people associate with such experiences. The trouble is that the following is extremely plausible:

**Extremely Plausible Thesis** – It’s not possible to teach a congenitally blind person what it’s like to see red by telling them about the causal role ordinary people associate with red experiences.
Extremely Plausible Thesis entails the falsity of no-compromise response to the knowledge argument. I am inclined to think that no causal structure information can teach you what it’s like to see red, but this thesis is more than we need to respond to the no-compromise response.

Perhaps it’s wrong to interpret Dennett as someone who thinks there are truths concerning what it’s like to see red. Much of what Dennett has written makes it natural to associate him with an eliminativism about phenomenal consciousness, and so maybe we should interpret him as thinking that there is nothing that it’s like to see red, and hence that there’s nothing that Mary learns when she sees red. However, I have already said in chapter 1 that I’m taking realism about phenomenal consciousness – as expressed by the Consciousness Constraint outlined in chapter 1 – for granted. I do not claim that I know for certain that there is something that it’s like to see red. But I take myself to be more justified in believing that there’s something that it’s like to see red – I contemplate this as I stare at bright red book cover in front of me – than I am in believing that there is an external world.

3.2.2 Non-Propositional Knowledge Responses

A less extreme response to the knowledge argument concedes that Mary learns something new, but holds that this new knowledge falls short of propositional knowledge. This response comes in two forms: the Ability Hypothesis and the Acquaintance Hypothesis.

According to the Ability Hypothesis, defended by Lawrence Nemirow and David Lewis, the knowledge Mary gains is a distinctive kind of know-how: a set of abilities she lacked when she was stuck in the black and white room. These abilities include the capacity to imagine and remember to see red, and to categorise external objects as red. Crucially, there is no new information Mary learns when she experiences red: she already knew all the propositional truths about red experiences in her black and white room. Compare: when you learn to wiggle your ears you don’t thereby learn any new facts about the world.

According to the Acquaintance Hypothesis, defended by Earl Conee, Mary gains knowledge by acquaintance, where acquaintance ‘requires the person to be familiar with the known entity in the most direct way that it is possible for a person to be aware of that thing.’\(^5\) You could know all sorts of truths about the great Liverpudlian comedian Ken Dodd, but until you’ve actually met him you don’t ‘know’ him, in the sense of knowledge by acquaintance. Similarly, Mary in her black and white room knows all the propositional truths there are to know about red experiences, but she doesn’t ‘know’ red experiences, in the sense of knowledge by acquaintance, until she actually has one.\(^6\)

No-Propositional Knowledge Responses suffer from a problem analogous to the ‘Frege-Geach problem’ for expressivist views in metaethics.\(^7\) Talk of what it’s like to see red can feature in a truth-preserving argument. For example, when Mary first learns what it’s like to see red, she might have had the thought common to many philosophically-inclined children that perhaps what it’s like for her to see red is different to what it’s like for others to see red. In entertaining these skeptical wonderings, she might consider the force of the following argument:

1. If this is what it’s like for me to see red, then this must be what it’s like for every other human being to see red.
2. Geoffrey is a human being.
3. Therefore, this is what it’s like for Geoffrey to see red.

\(^5\)Conee 1994: 144.

\(^6\)I am interpreting the Knowledge by Acquaintance Response as involving a notion of acquaintance such that being acquainted with experiences does not involve gaining any information about the experience. Many anti-physicalists believe in a kind of acquaintance that provides knowledge of the metaphysical nature of experiences, but a commitment to this kind of acquaintance strengthens the Knowledge Argument rather than constituting a response to it.

\(^7\)Geach 1965.
Whether or not this argument is sound it is certainly valid, in the sense that if its premises are true the truth of its conclusion follows. The only way we can make sense of this is by supposing that the phrase ‘This is what it’s like to see red’ expresses a truth-evaluable proposition. If so, and if that sentence also expresses what Mary learns when she sees red for the first time, it follows that Mary gains new propositional knowledge, rather than mere know-how or knowledge by acquaintance.⁸

There is another less technical concern with the Non-Propositional Knowledge Responses. It is evident that we can wonder about, be curious about, what it’s like to have certain experiences. I can be desperately curious about what it’s like to taste marmite, and my curiosity can be satisfied when I finally taste it, ‘Ah, so that’s what it’s like.’ Before Mary leaves her black and white room she might be desperately curious about what it’s like to see red, and be overjoyed when her curiosity is finally satisfied. We can’t make sense of this without supposing that there is some information about which we are curious, the having of which satisfies the curiosity. But according to the Non-Propositional Responses there is no information we gain in having experiences.

Of course one can wonder what it’s like to be able to ride to be able to ride a bike, or personally to know Ken Dodd. But plausibly such wonderings do not strictly speaking concern the having of an ability or the knowing of a person, but rather concern the experiences typically involved in exercising that ability or socialising with the person in question. I want to know what it’s like to ride along with the wind in my hair, or to actually feel Ken’s wrinkly hand gripping mine. If, as proponents of the No-Propositional Response allege, there is no new information one learns when one has an experience for the first time, then we cannot make sense of the having of an experience satisfying a curiosity.

Both because the sentence ‘This is what it’s like to see red’ can be embedded in the premise and conclusion of a truth-preserving argument, and because one can be curious about what it’s like to see red, the No Propositional Knowledge Responses are inadequate.

⁸ This argument is from Loar 1990/97.
3.2.3 The New Truth/Old Property Response

Any adequate response to the Knowledge Argument must concede that Mary gains new propositional knowledge when she sees red for the first time. This is a significant conclusion, for it follows that there is an *epistemic gap* between the pure physical and the experiential:

*The Epistemic Gap* – For any pure physical truth (i.e. a physical truth specified in a purely mathematico-nomic vocabulary) $C$, and any phenomenal truth (i.e. a truth concerning the instantiation of a conscious state, conceived of in terms of what it’s like to have it) $Q$, it’s not the case that $C$ a priori entails $Q$ (i.e. it’s not the case that one can rule out $<C\&\sim Q>$ a priori).

From this point onwards I will take this epistemic gap between the pure physical and the experiential to have been established. However, as I remarked at the start of this chapter, this significant *epistemological* conclusion is not enough to refute the *metaphysical* doctrine of physicalism. Considering the possibility of the New Truth/Old Property Response to the knowledge argument reveals why this is so.

Post-liberation Mary is able to think about red experiences in a way she wasn’t able to when she was trapped in the black and white room. She can bring to mind a red experience and think about it in terms of what it’s like to have it. You simply can’t think about red experiences in this way if you’ve never before had a red experience.\(^9\) Such concepts, of the

\[^9\text{It may of course be a contingent fact about humans that they get to possess a phenomenal concept of experience } E \text{ only by having experience } E. \text{ This does not contradict the crucial point Mary gains new propositional knowledge, and its implication that Mary gains a new concept. The epistemic gap can be further supported with Martina Nida-Rümelin’s (1996, 1998) thought experiment about Marianna. Unlike Mary, Marianna gets to know colours through arbitrarily coloured objects. Even though she has phenomenal concepts of various colour experiences, she will be unable to work out the relationship between these colour experiential states and pure physical states. Hence, even when one has the relevant phenomenal concepts, one cannot move a priori from pure physical truths to experiential truths.}\]
kind one deploys when one thinks about a conscious state in terms of what it’s like to have it, have become known as ‘phenomenal concepts’. When Mary sees red for the first time she gains a phenomenal concept of red experience.

However, from the mere fact that post-liberation Mary has a new concept, a new way of thinking about a feature of reality, it does not follow that the feature of reality she thinks about when she employs that concept is also new. Consider the following analogy. Suppose Mary has a quite ordinary child, Frank. Mary is keen on encouraging Frank to take an interest in astronomy, and points out the Morning Star to him when he is five year’s old. Frank subsequently thinks a lot about the Morning Star, and enjoys spotting in the sky. Now suppose that when he is eight year’s old Frank hears about Venus, and upon questioning his mother is told that the Morning Star is Venus. Eight year old Frank has picked up a new concept, a new way of thinking something in the world. But he did not thereby learn about some new thing in reality: the concept is new, but what it refers to is something Frank already knew about, namely the Morning Star.

The New Truth/Old Property Response to the knowledge argument makes an analogous claim about Mary. Pre-liberation Mary knew about brain state X, the brain state involved in seeing red. Post-liberation she gains a new phenomenal concept, a new way of thinking introspectively about a certain feature of reality. But her new phenomenal concept refers to something she already knew about pre-liberation, namely brain state X. ‘What it’s like to see red’ and ‘brain state X’ – just like the Morning Star and Venus – refer to one and the same feature of reality.

Thus, Mary gains brand new full-fledged propositional knowledge when she sees red for the first time. But that knowledge consists of a different way of thinking about physical properties she already knew about: she learns that this thing (brain state X picked out under a phenomenal concept) is identical with that thing (brain state X picked out with a pure

---

10 The term ‘phenomenal concepts’ comes from Chalmers 1996, but use of the term has subsequently become widespread on both sides of the debate.
physical concept. The knowledge argument refutes physicalism only if Mary learns about new properties when she leaves the room: these would have to be non-physical properties, given that she already knew about all the physical properties in the room. If what Mary learns is just new ways of thinking about properties she already knew about, then physicalism is safe.\footnote{This is the most popular response of type-B physicalists, whom we shall discuss in more detail in the next chapter. I will leave referencing them till then.}

Some argue that whenever one learns a new way of thinking about something, one thereby discovers a new property of that thing. It might be supposed, for example, that when Frank goes from thinking about the Morning Star as the Morning Star, to thinking about the Morning Star as ‘Venus’, he thereby learns a new property of the Morning Star: the property of being the second planet from the Sun. If this is always the case, then in gaining a new concept of red experiences, Mary would thereby learn about a new property of red experiences, and the threat to physicalism would return (For given that Mary didn’t know about this property in the black and white room, we are led to the conclusion that that property is non-physical).

However, it is not obviously true that possession of a new concept always comes along with knowledge of a new property. To return to the above example, it is possible to possess the concept of Venus without knowing that it’s the second planet from the Sun (I had to Google it to make sure). And it’s conceptually coherent to suppose that Venus is not a planet at all. We can imagine discovering that in fact Venus is a spaceship created by aliens to spy on us. This hypothesis may not be empirically very plausible, but it is not incoherent, from which it follows that it is not a priori that Venus is a planet. It is possible that in picking up the concept ‘Venus’ Frank learns nothing new about the Morning Star; he simply gains a new label for it. Analogously, according to the Old Fact/New Property Response, when Mary leaves her room she simply gains a new label for a feature of the world she already knew about.
Some would still want to push the point: there must be something new Frank learns about the Morning Star when he learns that it’s Venus. There must be some positive information associated with the concept ‘Venus’, in terms of which the concept characterises its referent. Likewise, one might be tempted to think, there must be some positive information associated with the phenomenal concept of red experience, in terms of which the concept characterises the experience of red. But at this point we can see that the knowledge argument relies on quite substantive claims about the workings of our concepts: that each and every concept picks out its referent by latching onto some property of it, some property which is a priori accessible to the concept user. We will be discussing these matters in more detail in the next chapter. For now we can merely note that nothing in the story of black and white Mary justifies this highly contentious meta-semantic assumption.

3.3 Transparency and opacity: The moral of the story

It is useful to think about the Knowledge Argument in terms of the distinction between transparent concepts and opaque concepts introduced in the first chapter (which is closely connected to the notion of metaphysical analysis discussed sections 2.1.3 and 2.2.2):

- **Transparent Concept** – A concept C of entity E is transparent just in case C reveals the nature of E (i.e. what it is for E to be part of reality is a priori accessible for someone possessing C, in virtue of possessing C), e.g. <sphericity>, <party>.

- **Opaque Concept** – A concept C referring to entity E is opaque just in case C reveals little or nothing of the essence of P. e.g. <water>, <gold>.

---

12 As I am defining terms, P can be a priori accessible for X even if X is not intelligent enough to access P; intuitively the thought would be that if X had sufficient rational powers X would be able to access P. If we are worried about how to define idealised rational powers, we can define transparency as follows: a concept C referring to entity E is transparent just in case there is a possible world in which someone works out the essence of E a priori, in virtue of possessing C, and without any empirical information other than what is required in order to possess C.
It is extremely plausible that Mary gains a new concept when she sees red for the first time, a new way of thinking about red experiences. Suppose that that concept is opaque: it reveals nothing about the nature of red experiences. In that case, Mary will gain new propositional knowledge post-liberation, but that new propositional knowledge is no threat to physicalism, as all she has gained is a new way of pointing to something in the world, and what she is pointing at may very well turn out to be one of the physical states she already knew about in the black and white room.

Suppose in contrast that Mary’s new concept is transparent: it reveals the essence of red experiences. Now pure physicalism looks to be in trouble. For in this case Mary’s new knowledge is knowledge of the nature of red experiences, but if pure physicalism is true she already knew the complete nature of red experiences in knowing the pure physical truths, and hence there ought to be nothing more she can learn about their nature.\(^{13}\)

I will not here dwell too much on this conclusion, as this will be a recurring theme in the first of half of the book: what is really at the heart of the anti-physicalist case is a commitment to *Phenomenal Transparency*, the thesis that phenomenal concepts are transparent. In the context of the knowledge argument, what Mary learns is a worry for the physicalist only if the new phenomenal concept she gains reveals something of the essential nature of colour experiences. The problem is that the knowledge argument itself does not have the resources to establish that phenomenal concepts are transparent, without which Mary’s knowledge is no threat to physicalism. For this reason the knowledge argument, in and of itself, does not refute physicalism.

\[^{13}\text{There are all sorts of further complexities that we will deal with in chapters 4 and 5, e.g. I am assuming in this brief argument the falsity of Dual Carving (see section 5.7).}\]