The Knowledge Argument Against Physicalism

The knowledge argument is one of the main challenges to physicalism, the doctrine that the world is entirely physical. The argument begins with the claim that there are truths about consciousness that cannot be deduced from the complete physical truth. For example, Frank Jackson’s Mary learns all the physical truths from within a black-and-white room. Then she leaves the room, sees a red tomato for the first time, and learns new truths—new phenomenal truths about what it is like to see red. The arguer infers that, contrary to physicalism, the complete physical truth is not the whole truth. The physical truth does not determine or metaphysically necessitate the whole truth about the world.

This article discusses that argument’s structure, compares Jackson’s version with others, compares the knowledge argument with other anti-physicalist arguments, and summarizes the main lines of response. Eight controversial assumptions are identified. These are the assumptions that:

- the notion of the physical is coherent;
- the complete physical truth is accessible to the pre-release Mary;
- upon leaving the room, she learns something;
- the kind of knowledge she acquires upon leaving the room is informational knowledge, rather than ability knowledge, acquaintance knowledge, or something else;
- she gains new information, rather than old information represented in a new way;
- if the complete-knowledge claim and the learning claim are true, then what Mary learns when she leaves the room cannot be a priori deduced (deduced by reason alone, without empirical investigation) from the complete physical truth.
- if there are phenomenal truths that cannot be a priori deduced from the complete physical truth, then the complete physical truth does not metaphysically necessitate those phenomenal truths;
- the knowledge argument and epiphenomenalism are consistent.
Various criticisms and defenses of these assumptions are discussed.

Table of Contents

1. Introduction
2. The Knowledge Intuition and the Inference to Physicalism’s Falsity
3. Related Arguments
4. More Physicalist Responses
5. Non Physicalist Responses
6. Other Responses
7. Jackson’s Retraction
8. Summary of Assumptions and Criticisms
9. References and Further Reading

1. Introduction

The knowledge argument aims to refute physicalism, the doctrine that the world is entirely physical. Physicalism (also known as materialism) is widely accepted in contemporary philosophy. But some doubt that phenomenal consciousness—experience, the subjective aspect of the mind—is physical. The knowledge argument articulates one of the main forms this doubt has taken.

Frank Jackson gives the argument its classic statement (in Jackson 1982 and Jackson 1986). He formulates the argument in terms of Mary, the super-scientist. Her story takes place in the future, when all physical facts have been discovered. These include “everything in completed physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this, including of course functional roles” (Jackson 1982, p. 51). She learns all this by watching lectures on a monochromatic television monitor. But she spends her life in a black-and-white room and has no color experiences. Then she leaves the room and sees colors for the first time. Based on this case, Jackson argues roughly as follows. If physicalism were true, then Mary would know everything about human color vision before leaving the room. But intuitively, it would seem that she learns something new when she leaves. She learns what it’s like to see colors, that is, she learns about *qualia*, the properties that characterize what it’s like. Her new phenomenal knowledge includes knowledge of truths. Therefore, physicalism is false.

In the late 1990’s, Jackson changed his mind: he now defends physicalism and rejects the knowledge argument. But others defend the argument, and even those who reject it often disagree about where it goes awry. The knowledge argument has inspired a voluminous literature, which contains insights about consciousness, knowledge, the limits of third-person science, and the nature of the physical. It is also discussed in non philosophical works, including a book by E. O. Wilson (1998), a work of fiction (Lodge 2001), and a T.V. series (*Brainspotting*). This article discusses the argu-
ment’s structure, compares Jackson’s version with others, compares the knowledge argument with other anti-physicalist arguments, and summarizes the main lines of response.

2. The Knowledge Intuition and the Inference to Physicalism’s Falsity

The knowledge argument has two parts. One says that physical knowledge is not sufficient for phenomenal knowledge. Call this the knowledge intuition (Stoljar and Nagasawa, 2004). The other says that the knowledge intuition entails the falsity of physicalism.

Thus described, the knowledge argument is not new with Jackson. Locke and other 18th Century British empiricists discussed the knowledge intuition. C. D. Broad gave a version of the knowledge argument in 1925. And other versions appear in more recent writings, such as Thomas Nagel’s 1974 “What is it Like to be a Bat?” What is distinctive about Jackson’s contribution?

Daniel Stoljar and Yujin Nagasawa (2004) answer this question in their introduction to a volume of essays on the knowledge argument. As they say, Jackson contributes at least two main ideas: his Mary example illustrates the knowledge intuition better than previous attempts; and he provides distinctive reasons for inferring physicalism’s falsity from the intuition. Let us take these points in order.

The Mary case divides the knowledge intuition into three claims:

- The complete-knowledge claim: before leaving the room, Mary knows everything physical.
- The learning claim: upon leaving, she learns something.
- The non-deducibility claim: if the complete-knowledge claim and the learning claim are true, then what Mary learns when she leaves the room cannot be a priori deduced (deduced by reason alone, without empirical investigation) from the complete physical truth.

Physicalists may deny the knowledge intuition. But the Mary case suggests that doing so requires rejecting the complete-knowledge claim, the learning claim, or the non-deducibility claim.

The cases discussed by Broad, Nagel, and others do not deliver this result. Consider, for example, Broad’s “mathematical archangel,” a logically omniscient creature who knows all the physical truths about various chemical compounds. Broad calls these truths “mechanistic” instead of “physical,” but the point is the same. On his view, the archangel would know all such truths but still lack phenomenal knowledge concerning, for example, “the peculiar smell of ammonia.” And Broad infers that physicalism (“mechanism”) is false. But what if the physicalist denies that the archangel would lack the relevant phenomenal knowledge? We appear to be at an impasse. By contrast, if the physicalist claims that, while in the room, Mary knows what it’s like to see colors, he must explain
why she seems to acquire this knowledge when she leaves. The Mary case breaks the deadlock in favor of the knowledge intuition. Other illustrations of the intuition that precede Jackson’s have further drawbacks. For example, Nagel’s claim that humans cannot imagine what it’s like to be a bat raises distracting issues about the limits of human imagination, about which physicalism carries no obvious commitments. Mary’s fame is just.

To explain the second of Jackson’s distinctive contributions, it will be useful to explain some terminology and abbreviations. First, there is the distinction between the \textit{a priori} and the \textit{a posteriori}. A priori truths are those that are justifiable by reason alone, without empirical investigation. Logical truths provide clear examples. For example, one can figure out without empirical investigation that the following claim is true: if Socrates is mortal, then either Socrates is mortal or Socrates is fat. Compare the claim that Socrates is mortal. While we believe the latter claim to be true, reason alone does not justify this belief. Instead, we rely on experience—empirical investigation. So, while it is \textit{a priori} that if Socrates is mortal, then either Socrates is mortal or Socrates is fat, it is \textit{a posteriori} that Socrates is mortal. We may also speak of truths that are \textit{a priori} deducible from other truths. For example, although “Socrates is mortal” is \textit{a posteriori}, that same truth is \textit{a priori} deducible from two other truths: “All men are mortal” and “Socrates is a man.” In other words, the latter two truths, taken together, \textit{a priori} entail that Socrates is mortal.

Second, there is the notion of \textit{metaphysically necessary} truths. A \textit{necessary} truth is a truth that could not have failed to be the case. Logical truths again provide clear examples: “Either Socrates is mortal or it is not the case that Socrates is moral” is usually regarded as necessary. Contrast that truth with “Socrates is mortal.” The latter is not necessary. Truths that are not necessary are also known as \textit{contingent}. Philosophers often distinguish between different strengths or kinds of necessity. For example, there is arguably a sense in which it is a necessary truth that pigs cannot fly like birds. But if the laws of nature were different, then perhaps pigs would be able to fly like birds. So, perhaps it is not metaphysically impossible that pigs should be able to fly like birds. A \textit{metaphysically necessary} truth is a truth that is necessary in the strictest possible sense: a truth that holds not just because of contingent laws of nature. Saul Kripke (1972) famously argues that there are metaphysically necessary truths that are not truths of pure logic. Indeed, he argues that there are metaphysically necessary truths that are not \textit{a priori}. For example, on his view, \textit{that water is H2O} is metaphysically necessary but \textit{a posteriori}. He recognizes that there could have been substances that \textit{resemble} water—substances that share water’s superficial qualities, such as its taste and visual appearance—but with a different molecular structure. But, he argues, these substances would not be water.

Third, let us introduce some abbreviations. On Jackson’s version of the knowledge argument, the assumption that Mary knows the complete physical truth about the world does not guarantee that she will be able to figure out the complete truth about human color vision. His reasoning involves the idea of the complete physical truth. Call the complete physical truth \( P \). \( P \) can be seen as a long
conjunction of all the particular physical truths, which, according to Jackson, Mary learns from watching science lectures. What about the truths that, according to Jackson, Mary does not learn until she leaves the room? Those would be included in the psychological truths about the world. Call the complete psychological truth Q. Finally, consider what Stoljar and Nagasawa call “the psychophysical conditional”: if P then Q, where P is the complete physical truth and Q is the complete psychological truth. As we will see, part of Jackson’s reasoning can be understood in terms of his view about the psychophysical conditional.

We are now in a position to state the second of Jackson’s distinctive contributions to the discussion of the knowledge argument. This contribution concerns his inference from the knowledge intuition to physicalism’s falsity. His inference assumes that if physicalism is true then the complete truth about human color vision is a priori deducible from the complete physical truth. But here a problem arises: why accept this assumption? Consider the psychophysical conditional, if P then Q (again, P is the complete physical truth and Q is the complete psychological truth). As Jackson conceives of physicalism, physicalism entails that the psychophysical conditional is a priori. If he is right, then all truths about color vision would be deducible from P (the complete physical truth). But here physicalists have a natural, obvious response: why not instead characterize physicalism as a Kripkean a posteriori necessity, akin to water is H2O? On this characterization, the psychophysical conditional is metaphysically necessary but not a priori.

In later work, Jackson criticizes this response. His argument is complex, but the basic idea is simple enough. In a 1995 “Postscript,” he reasons as follows. Consider the argument:

H2O covers most of the planet.
Therefore, water covers most of the planet.

The premise necessitates, but does not a priori entail, the conclusion. But, Jackson asks, why is there no a priori entailment? On his view, there is no such entailment because the argument’s premise gives us only part of the physical story. It is also part of the physical story that H2O does the other things that water does, that is, that H2O plays the water role. Playing the water role includes such things as being a substance that occupies oceans and lakes, looks clear to us, has little or no taste, is referred to as “water”, etc. So, let us add the following premise to the argument displayed above:

H2O plays the water role.

Now, says Jackson, the premises do a priori entail the conclusion. Moral: “a rich enough story about the H2O way things are does enable the a priori deduction of the water way things are” (Jackson 1995, p. 413). Likewise, physicalism entails that “knowing a rich enough story about the physical nature of our world is tantamount to knowing the psychological story about our world” (Jackson 1995, p. 414). But if physicalism is true, P should provide just that: a rich enough story.
Thus, Jackson concludes, physicalism entails the apriority of the psychophysical conditional after all.

Jackson’s argument is controversial. But in developing it, he fills an important lacuna in the knowledge argument and thereby improves on earlier versions. Others, too, have attempted to fill this lacuna. Most notably, David Chalmers (1996, 2003, 2004, and 2006a) has given sophisticated arguments to this end, which are partly inspired by Jackson’s argument.

3. Related Arguments

The knowledge argument is one of several ways to articulate the suspicion that phenomenal consciousness is not physical. Another common way of articulating the doubt is through the conceivability argument. This argument descends from René Descartes’ main argument for mind-body substance dualism. He argued that, since he can clearly and distinctly conceive of his mind without his body and his body without his mind, they can exist without each other and are therefore distinct substances.

Contemporary versions of the conceivability argument usually rely on thought experiments concerning qualia. One such thought experiment involves inverted qualia. It seems conceivable that there be an individual exactly like me, except he and I are red/green inverted. We are physically and functionally identical, but the color experiences he has when viewing a ripe tomato (in normal light, without special contact lenses, and so forth) resemble the color experiences I have when viewing a ripe zucchini, and vice versa. Such a person would be my inverted twin. Likewise, it seems conceivable that there be a world exactly like ours in all physical and functional respects but without phenomenal consciousness. Creatures that lack consciousness but are physically and functionally identical to ordinary human beings are called zombies. If it is conceivable that there be creatures such as my inverted twin or my zombie twin, then, the conceivability argument runs, this supports the metaphysical possibility of such creatures. And most agree that if such creatures are metaphysically possible, then phenomenal consciousness is neither physical nor functional: physicalism is false.

Yet another related argument is the explanatory argument. This argument begins with the premise that physicalist accounts explain only structure (such as spatiotemporal structure) and function (such as causal role). Then it is argued that explaining structure and function does not suffice to explain consciousness, and so physicalist accounts are explanatorily inadequate.

As Chalmers (2003) notes, the knowledge argument, the conceivability argument, and the explanatory argument can be seen as instances of a general, three-step argument. The first step is to establish an epistemic gap between the physical and phenomenal domains. In the case of the knowledge argument, the gap is often put in terms of a priori deducibility: there are phenomenal truths that
cannot be \textit{a priori} deduced from physical truths. In the case of the conceivability argument, the gap is put in terms of conceivability: it is conceivable that there be inverted qualia or zombies. And in the case of the explanatory argument, the point is put in terms of an explanatory gap. After establishing an epistemic gap, these arguments take a second step and infer a corresponding \textit{metaphysical gap}: a gap in the world, not just in our epistemic relation to it. The knowledge argument infers a difference in type of fact. The conceivability argument infers the metaphysical possibility of inverted qualia or zombies. And the explanatory argument infers that there are phenomena that cannot be physically explained. As a third step, all three results appear to conflict with physicalism. There are important differences among the arguments, and it is not obvious that they stand or fall together. Nevertheless, it is worth noting that they follow a single abstract pattern.

\section*{4. More Physicalist Responses}

Most physicalist responses to the knowledge argument fall into three categories: those that reject the inference to physicalism’s falsity and thus deny the metaphysical gap; those that reject the knowledge intuition and thus deny the epistemic gap; and those that derive an absurdity from Jackson’s reasoning.

We have already noted one way of rejecting the inference from the knowledge intuition to physicalism’s falsity: one could defend a version of physicalism on which the psychophysical conditional is necessary but not \textit{a priori}. There are other ways of rejecting the inference. One is to reject the assumption that phenomenal knowledge is propositional knowledge—knowledge of truths or information. That is, one could argue that the \textit{type} of knowledge Mary gains when she leaves the room is non propositional. The most popular version of this view is based on the ability hypothesis, the claim that to know what it’s like is to possess certain \textit{abilities}, such as the ability to imagine, recognize, and remember experiences. On this view, Mary’s learning consists in her acquiring abilities rather than learning truths. As the view is sometimes put, she gains \textit{know-how}, not \textit{knowledge-that}. There are other versions, including the view that upon leaving the room Mary acquires only non propositional acquaintance knowledge ( Conee 1994, Bigelow and Pargetter 1990). On this version, her learning consists, not in acquiring information or abilities, but in becoming directly acquainted with the phenomenal character of color experiences, in the way that one can become acquainted with a city by visiting it.

These views allow the physicalist to accept the knowledge intuition without facing objections that Jackson, Chalmers, and others bring against \textit{a posteriori} physicalism. But other problems arise. Regarding the ability hypothesis, some doubt that Mary’s learning could consist only in acquiring abilities. Her new knowledge appears to have characteristic marks of propositional knowledge because its content can be embedded in conditionals such as “if seeing red is like this, then it is not like that” (Loar 1990/97). And some philosophers question the significance of the distinction be-
tween know-how and knowledge—that on which the strategy of the ability-hypothesis seems to rely (Alter 2001, Stanley and Williamson 2001).

The idea that Mary acquires only acquaintance knowledge has similar difficulties. It is not clear that all she acquires is acquaintance knowledge or that the requisite distinction between acquaintance knowledge and propositional knowledge is tenable. Also, there is a danger of trading on an ambiguity: sometimes “acquaintance” refers to knowledge, sometimes to experience. On the former, epistemic interpretation, it is unclear that Mary’s new “acquaintance knowledge” includes no factual component. And on the latter, experiential interpretation, the acquaintance hypothesis trivializes the learning claim: no one denies that when Mary leaves the room she has new experiences.

Another way to reject the inference to physicalism’s falsity is to argue that Mary’s learning consists in acquiring new ways to represent facts she knew before leaving the room (Loar 1990, 1997, Lycan 1996, Horgan 1984, McMullen 1985, Pereboom 1994, Tye 2002). This view is often combined with an appeal to a posteriori necessity (see section 2 above). But it need not be: one could argue that while the psychophysical conditional is a priori knowable by those who possess the relevant phenomenal concepts, Mary lacks those concepts before leaving the room. The main challenge for this view concerns the status of her new concepts. It is not enough to say that she gains some new concept or other: her conceptual gain must explain her gain in knowledge. The concern is that any concepts adequate to the task—such as the concept having an experience with phenomenal feel f—might incorporate a non physical component (Chalmers 2006b).

Philosophers have also devised ways to reject the knowledge intuition. Some believe that intuitions based on hypothetical cases should be given little or no weight. Also, specific strategies for rejecting the knowledge intuition have been developed. One is to reject the learning claim: to argue that on reflection Mary does not learn anything when she leaves the room. Some defend this position by arguing that we simply underestimate the power of complete physical knowledge. Suppose we try to fool Mary by greeting her when she leaves the room with a blue banana. Would she be fooled into thinking that seeing yellow is what we would describe as seeing blue? Not necessarily. She could use a brain scanner (perhaps a descendent of a PET device) to examine her own brain processes. She would notice that her brain processes correspond to people having blue experiences, and thereby evade our trap. Maybe our intuition that she learns something fails to take this sort of consideration into account (Dennett 1981, 2006). But other philosophers doubt that the intuition derives from any such error.

Another way to reject the knowledge intuition is to challenge the complete-knowledge claim: to argue that not all physical facts about seeing colors can be learned by watching black-and-white lectures. On this view, a fact might be physical but not discursively learnable. How could this be?
Some (for example, Horgan, 1984) use “physical” broadly, so that the physical truths include high-level truths necessitated by the microphysical truths. These physicalists argue that phenomenal truths are themselves high-level physical truths, and that it is question-begging to assume that Mary knows all the physical truths simply because she watches lectures on chemistry, physics, etc. Chalmers (2004, 2006a) suggests a natural response to this move: use “physical” narrowly, so that the physical truths include only the microphysical truths (or those plus the truths in chemistry or some other specified domains). It is harder to deny that such truths would be accessible to the pre-release Mary. Of course, this entails that high-level biological truths, for example, will count as non physical, and thus the existence of non physical truths will not itself defeat physicalism. But if Jackson’s reasoning is sound, then there are phenomenal truths that are not metaphysically necessitated by the narrowly physical truths—and that result would defeat physicalism.

On another version of the view that the complete-knowledge claim is false, Mary’s science lectures allow her to deduce the truths involving structural-dynamical properties of physical phenomena, but not their intrinsic properties. The knowledge argument does not appear to refute this view. If this view can reasonably be called a physicalist view, then there is at least one version of physicalism that the knowledge argument appears to leave unchallenged. However, it is unclear that this is a significant deficiency. Arguably, on the view in question, consciousness (or protoconsciousness) is a fundamental feature of the universe—or at least no less fundamental than the properties describable in the language of physics, chemistry, etc. That sounds like the sort of view the knowledge argument should be used to establish, not refute. (It is a form of neutral monism; see next section.)

5. Non Physicalist Responses

If we accept the knowledge argument, then how should we understand the relationship between consciousness and the physical world? Jackson (1982) defends epiphenomenalism, on which phenomenal properties or *qualia* are caused by but do not cause physical phenomena. But epiphenomenalism is only one non physicalist view that the knowledge argument leaves open. Others include interactionism, parallelism, and idealism. These views agree that consciousness is not reducible to the physical, but disagree over how the two interact causally. On interactionist dualism, consciousness affects the physical world and vice versa. On parallelism, physical events and events of consciousness run in parallel but do not affect each other. On idealism, there are only conscious phenomena. The knowledge argument also leaves open neutral monism, the view that phenomenal properties (or protophenomenal properties) are the categorical, intrinsic bases of physical properties, which are at bottom dispositional and relational. This view might or might not be considered a version of physicalism, depending on whether the intrinsic nature of physical properties is considered physical.

All of these views have significant costs and benefits. For example, interactionist dualism is commonsensical but hard to reconcile with the popular view that the physical world is causally closed,
that is, the view that every physical event has a sufficient *physical* cause. To take another example: epiphenomenalism preserves causal closure but seems to conflict with the widespread naturalistic assumption that consciousness is an integrated part of the natural world. Accepting the knowledge argument forces philosophers to weigh such costs and benefits and develop new, non physicalist accounts.

Historically, epiphenomenalism is associated with Huxley (1874), interactionist dualism with Descartes (1641), parallelism with Leibniz (1714), idealism with Berkeley (1713), and neutral monism with Russell (1927). For more recent versions, see Jackson (1982) and Robinson (1982b, 1988) for epiphenomenalism; see Popper and Eccles (1977), Hart (1988), Foster (1991), and Hodgson (1991) for interactionist dualism, see Rosenberg (2004) for neutral monism; and see Adams (forthcoming) for idealism. There are no recent defenses of parallelism.

6. Other Responses

Some claim that Jackson’s position is internally inconsistent (Watkins 1989, Campbell 2003). The argument runs roughly as follows. On the knowledge argument, Mary acquires knowledge when she leaves the room because she has states with new qualia. But this is impossible if, as Jackson (1982) suggests, epiphenomenalism is true: on epiphenomenalism, qualia are causally inefficacious; so, how can qualia produce an increase in knowledge? So, Jackson cannot consistently maintain both epiphenomenalism and the learning claim.

However, the sort of epiphenomenalism Jackson defends implies, not that phenomenal features are inefficacious, but only that they have no effects on *physical* phenomena. He might therefore reply that phenomenal knowledge is not a physical phenomenon, and thus qualia may indeed cause Mary to acquire it. Also, he can reasonably complain that the objection assumes a causal theory of knowledge that is not appropriate for phenomenal knowledge.

Despite the availability of these replies, there is a serious problem in the vicinity of the inconsistency objection. We should expect physical or functional explanations of our judgments about qualia. But if the knowledge argument is sound, then qualia would seem to be explanatorily irrelevant to these judgments—including the judgment that qualia cannot be explained in physical or functional terms. This is what David Chalmers calls “the paradox of phenomenal judgment” (Chalmers 1996, chapter 5). It appears to be a real problem, which arises for any non physicalist theory of consciousness.

Another important response to the knowledge argument should be noted. The argument seems to assume that “physical” has a clear meaning. But whether this notion can be adequately defined is not obvious. One problem is “Hempel’s dilemma” (Hempel 1980, Montero 1999). Arguably, we should not define the physical in terms of current physics, because current physics will be extended
and presumably revised in substantial ways. We could define it in terms of ideal physics. But who knows what ideal physics will look like? Future physics may involve novel concepts that we cannot begin to imagine. If “physical” is defined in terms of such unknown concepts, then how can we judge whether Mary could learn all the physical facts from black-and-white lectures? And how else should we define the notion except by appeal to (current or ideal) physics?

Some take such considerations to show that the debate over whether consciousness is physical is misguided or meaningless (Chomsky 1980, 1988, Crane and Mellor 1990, Montero 1999). But the difficulty may be surmountable. On one view, ideal physics will not be wholly unrecognizable: like today’s physics, it will be concerned entirely with structure and dynamics. And one may be able to argue that any structural/dynamical properties can in principle be imparted by black-and-white lectures.

7. Jackson’s Retraction

As we noted earlier, Jackson (1998, 2003, 2006) has come to embrace physicalism and reject the knowledge argument. More specifically, he rejects the claim that Mary learns new truths when she leaves the room. He argues that this claim derives from a mistaken conception of sensory experience—a conception that he thinks should be replaced with representationalism, the view that phenomenal states are representational states. Interestingly, he combines this view with the ability hypothesis. He writes,

Those who resist accounts in terms of ability acquisition tend to say things like “Mary acquires a new piece of propositional knowledge, namely, that seeing red is like this”, but for the representationalist there is nothing suitable to be the referent of the demonstrative.

We have ended up agreeing with Laurence Nemirow and David Lewis [the authors of the ability-hypothesis strategy] on what happens to Mary on her release. But, for the life of me, I cannot see how we could have known they were right without going via representationalism. (Jackson 2003, p. 439)

It is unclear why Jackson’s representationalism leads him to embrace the ability hypothesis. Despite his commitments to physicalism and the apriority of the psychophysical conditional, he has other options. For example, instead of explaining Mary’s epistemic progress in terms of newly acquired abilities, he might argue that her “progress” is an illusion; in other words, he might reject the learning claim. Moreover, it may be possible to formulate a representationalist version of the knowledge argument that inherits the force of the original (Alter, 2006).

8. Summary of Assumptions and Criticisms

As we have seen, the knowledge argument depends on several controversial assumptions. It will be useful to summarize some of these assumptions and some criticisms of them. I will also mention
some sources for relevant arguments.

**Assumption 1:** The coherence of the notion of the physical: physicalism is a substantive doctrine with non trivial content.

Criticism 1: The notion of the physical is not well defined, and there is no substantive issue of whether physicalism is true (Chomsky 1980, 1988, Crane and Mellor 1990, Montero 1999). For replies, see Chalmers (1996, 2004) and Stoljar (2000).

**Assumption 2:** The complete-knowledge claim (“truths” version): before leaving the room, Mary knows all physical truths.

Criticism 2a: Pre-release Mary does not know all the physical truths, because high-level physical truths cannot in general be a priori deduced from low-level physical truths (Horgan 1984, van Gulick 2004, Block and Stalnaker 1999). For replies, see Chalmers (2004, 2006a) and Chalmers and Jackson (2001).

Criticism 2b: Pre-release Mary does not know all the physical truths, because truths about the intrinsic properties of physical phenomena cannot be discursively learned (Alter 1998, Stoljar 2000). For replies, see Chalmers (1996, 2003, 2004, 2006a)

**Assumption 3:** The learning claim: upon leaving the room, Mary learns something.


Criticism 3b: We think Mary learns something because we fail to recognize that phenomenal properties are just representational properties (Jackson 2003). For a reply, see Alter (2006); and for a counter-reply, see Jackson (2006).

Criticism 3c: Mary gains only unjustified beliefs (Beisecker 2000).

**Assumption 4:** The non-deducibility claim: if Mary learns new phenomenal truths when she leaves the room, then those truths cannot be a priori deduced from the complete physical truth.

Criticism 4: Mary cannot deduce certain phenomenal truths from the complete physical truth only because she lacks the relevant concepts, such as the concept of phenomenal redness. Thus, even though Mary cannot deduce Q from P, the psychophysical conditional is a priori for those who have
Assumption 5: The propositional-knowledge claim: the kind of knowledge Mary gains upon leaving the room is propositional or factual—knowledge of information or truths.


Criticism 5c: Mary gains non propositional knowledge that does not fit easily into folk categories (Churchland 1985, 1989).

Assumption 6: The new-information claim: the information Mary gains upon leaving the room is genuinely new to her.


Assumption 7: The claim that the knowledge intuition entails non necessitation: if there are phenomenal truths that cannot be a priori deduced from the complete physical truth, then the complete physical truth does not metaphysically necessitate those phenomenal truths.

Criticism 7: Physicalism is an a posteriori necessity and is therefore compatible with the claim that the phenomenal truths are not deducible from the complete physical truth. For references, see the second version of criticism 6 above.

Assumption 8: The consistency claim: the knowledge argument and non physicalism are consistent.
Criticism 8: The assumption that Mary gains knowledge is inconsistent with epiphenomenalism (Watkins 1989, Campbell 2003). For replies, see Nagasawa (n.d.).

The knowledge argument rests on other assumptions. One is that if Mary gains new, non physical information, then there are non physical properties. Another is that if there are truths that are not metaphysically necessitated by the complete physical truth, then physicalism is false. But doubts about these assumptions may be terminological variants on doubts about assumptions 1-8.

Some critics combine elements of different criticisms. For example, Michael Pelczar’s (forthcoming) criticism appears to contain elements of the acquaintance hypothesis and the old-fact/new-representation view; Jackson both rejects the learning claim and endorses the ability hypothesis (Jackson 2003); and Robert van Gulick (2004) argues that the various physicalist criticisms of the knowledge argument can be seen as parts of a single, coherent reply. Those who endorse the knowledge argument (in addition to Jackson, before he changed his mind) include Robinson (1982a), Nida-Rümelin (1995), Chalmers (1996, 2004, 2006a), and Gertler (1999).

William Lycan (2003) writes, “Someday there will be no more articles written about the “Knowledge Argument”... That is beyond dispute. What is less certain is, how much sooner that day will come than the heat death of the universe” (Lycan 2003). At least for now, however, the knowledge argument continues to inspire fruitful reflection on the nature of consciousness and its relation to the natural world.

9. References and Further Reading


Berkeley, George. 1713. Three Dialogues between Hylas and Philonous.


Descartes, René. Meditations on First Philosophy. 1641.


Leibniz, G. 1714. The Monadology.


Nagasawa, Yujin, n.d. The “Most Powerful Reply” to the Knowledge Argument. ANU manuscript.


**Author Information**

Torin Alter  
Email: talter@as.ua.edu  
The University of Alabama  
U. S. A.