DOI: 10.36446/af.2023.477

WHERE DOES THE KNOWLEDGE ARGUMENT GO WRONG?

¿Dónde erra el Argumento del Conocimiento?

Fernando Rudy Hiller [a]
https://orcid.org/0000-0002-7977-1216
rudy@filosoficas.unam.mx

[a] Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México, Ciudad de México, México.

Abstract

In his well-known Knowledge Argument (KA) Frank Jackson attempted to show that physicalism is false by offering a case that allegedly showed that a complete physicalist description of the world leaves something crucial out, namely the phenomenal qualities of experience. Eventually Jackson himself retracted and claimed that the interesting task is to explain where and why intuition-pumping arguments against physicalism such as the KA go wrong. This is exactly the task that occupies this paper: to discuss and criticize three of the most important diagnoses of the KA's weak points and to offer my own view about the latter. Along the way, several important but often neglected features of the KA are expounded and clarified.

Key words: Knowledge Argument; Jackson; Physicalism; Qualia; Supervenience.

Resumen

Con su bien conocido Argumento del Conocimiento (*Knowledge Argument*, KA) Frank Jackson intentó establecer la falsedad del fisicalismo al ofrecer un caso que supuestamente mostraba que una descripción exhaustiva del mundo en términos fisicalistas dejaba fuera algo crucial, a saber, las cualidades fenoménicas de la experiencia. Eventualmente Jackson mismo se retractó y llegó a afirmar que la tarea interesante es explicar dónde y por qué erran los argumentos basados en intuiciones en contra del fisicalismo. Esta es precisamente la tarea que ocupa este trabajo: discutir y criticar tres de los más importantes diagnósticos acerca de los puntos débiles del KA y ofrecer mi propia posición al respecto. En el curso del mismo se exponen y clarifican distintas características centrales del KA que usualmente son pasadas por alto.

Palabras clave: Argumento del Conocimiento; Jackson; Fisicalismo; Qualia, Superveniencia.

1. Introduction

In his paper "Mind and Illusion" (2004), Frank Jackson disavows his earlier antiphysicalist sympathies and offers a rebuttal to his own Knowledge Argument (KA) against physicalism. He explains his change of heart thus:

> Most contemporary philosophers, when given a choice between going with science and going with intuitions, go with science. Although I once dissented from the majority, I have capitulated and now see the interesting issue as being where the arguments from the intuitions against physicalism—the arguments that seem so compelling—go wrong (Jackson, 2004, p. 421).

As Jackson's later physicalist self, I think that intuition-pumping arguments against physicalism *must* be wrong; the challenge is to say where and why. The literature about the KA is vast; my contribution to it consists, first, in offering a comprehensive presentation of the most important physicalist responses to the KA—Daniel Stoljar's two conceptions of the physical; the Ability Hypothesis by David Lewis, and the "Old Facts, New Modes" thesis by Brian Loar. Second, in evaluating the strength and weaknesses of each; and third, in showing where the crux of the debate really lies—*i.e.*, the notion of phenomenal information. While I don't offer here an entirely novel position, when it comes to a well-trodden but often confusing debate such as the one concerning the KA, the three contributions listed above are, I think, significant. The structure of the paper is as follows: in section 2, the KA is introduced along with some basic notions that help understand why it seems to pose such a strong challenge to physicalism; in sections 3 through 5 three physicalist responses to the argument are explored and criticized and my own view expounded; section 6 offers a brief conclusion.

2. The Knowledge Argument

The KA purports to show that truths about qualia—the phenomenal aspect of psychological states—cannot be deduced from physical truths, and hence that physicalism about the mind is false. The thought experiment on which the KA is based is well-known: Mary, a neuroscientist with exhaustive knowledge of the physical functioning of the brain and especially of color vision science, has spent her whole life locked up in a black and white room, so she lacks any color experience.

One day she is released from her prison and sees red for the first time. The crucial question is: Does Mary learn anything new about color vision or not? The KA suggests an affirmative answer: after being released, Mary does learn something new, namely, *what it is like* to see red. More importantly, she comes to realize that her previous knowledge about the visual perception of people *in general* (not just about her own[1]) was incomplete: there is some further fact about the color vision of human beings that she missed despite her complete physical information about it. Hence, the argument concludes, this further fact—the "what it is like" fact—cannot be a physical fact; but it is a fact about people nonetheless, so physicalism must be false.

Jackson's formal rendering of the KA in "What Mary Didn't Know" (1986/2004, p. 54) goes as follows:

> (1) Mary (before her release) knows everything physical there is to know about other people.
>
> (2) Mary (before her release) does not know everything there is to know about other people (because she *learns* something about them on her release).
>
> Therefore,
>
> (3) There are truths about other people (and herself) which escape the physicalist story.

The argument is deceptively simple and, in the face of it, very powerful. However, the hypothesis under which I will proceed is that it is mistaken; but where does the mistake lie? Different philosophers have located the mistake in different parts of the argument. Concerning the first premise, Daniel Stoljar (2001/2002) has denied that Mary knows everything physical about other people before being released. Concerning the second premise, David Lewis (1988 /2002) accepts that Mary learns something after being released but argues that what she learns is non-factual (and hence non-propositional) knowledge, which means that there is an equivocation in the use of the crucial term 'knowledge' at the

[1] The proviso is needed to avoid the objection that Mary was not ignorant about any fact about her *own* color experience, since she had none before being released. But if we include other people, it becomes clear (according to the KA) that she was ignorant of some fact about them when she was in her room thinking she knew all there was to know about color vision.

heart of the KA, since the first premise clearly refers to propositional knowledge. Finally, concerning the conclusion, Brian Loar (1997/2002) has claimed that, although Mary comes to know non-physical truths (or truths that cannot be couched in physicalist vocabulary) after being released, this fact does not threaten physicalism because what it shows is that phenomenal *concepts* cannot be reduced to physical ones, not that phenomenal *properties* are themselves independent of physical ones. (What the story of Mary shows, according to Loar, is that the identity of phenomenal and physical properties is a necessary a posteriori one. We will return to this below.) My own diagnosis is that the mistake lies in the second premise: Mary indeed changes when she sees red for the first time, but there are deep problems in characterizing this change as the acquisition of new knowledge. As we will see, the antiphysicalist lacks a convincing story of what this new knowledge comes to. Thus, while I am very sympathetic to Lewis' response to the KA, I do not find his defense of the Ability Hypothesis compelling (see section 4 below).

Before starting the analysis of the rebuttals to the KA, I will explain two central points for the discussion that follows: 1) what is physicalism and why the KA presents a direct threat to it; 2) what is the psychophysical conditional and why it must be—if physicalism is true—necessarily true and, according to Jackson, true a priori.

Physicalism and supervenience

In "A Definition of Physicalism", Philip Pettit (1993, p. 213) argues that physicalism is a thesis composed of two main ideas: 1) the materials the empirical world is made of are the ones identified by physics; 2) the empirical world is governed by the forces or regularities described by physics. Pettit takes for granted that the constitutive materials and the governing forces are those of microphysics. Each of the above ideas incorporates two basic claims: a) microphysical entities and regularities exist (physicalism is a realist position concerning the metaphysical status of scientific theories); b) microphysical entities constitute everything and microphysical regularities govern everything. The physicalist claim that is directly denied by the KA is the claim that everything in the empirical world is constituted by the microphysical entities postulated by physics.

Let us take a closer look at this claim. Pettit remarks that it incorporates a crucial *supervenience thesis*: "two microphysically composed entities cannot differ intrinsically without some difference of a microphysical kind" or, more straightforwardly, "No macrophysical

difference without a microphysical one" (1993, pp. 215-216). In other words, the macrophysical supervenes on the microphysical. This supervenience claim is precisely what the KA attempts to show is false: when Mary is released and sees red for the first time, she comes to know that the macrophysical world she sees is different from the macrophysical world described by microphysics in that the former incorporates qualia. But qualia (the "what it is like" of seeing red, for instance) is presumably a macrophysical feature of the empirical world, a feature that is not accompanied by any microphysical difference because, so far as microphysics is concerned, a world with qualia is identical (microphysically speaking) to a world without qualia. (This can be seen from the fact that nothing in Mary's exhaustive knowledge of microphysics predicted the existence of qualia.) Hence, the supervenience claim does not hold, and physicalism is false—or so the KA tries to show.

Another way of making the same point is Lewis' (1988/2002, p. 286): physicalism is the thesis that "any two possibilities that are just alike physically are just alike *simpliciter*." This implies that two physical duplicates cannot be differentiated by means of *physical information*: anything that you say about the one you must say about the other. In the case of Mary, we can say the following: Mary has exhaustive physical knowledge about the color vision of people; so she expects that once she is released, the knowledge she has will not be altered, because the people she will encounter outside of her room are physically alike to the people described in her physiology books—actually, they are the same. So let us say that people described in her books are one "possibility" in Lewis' sense and people outside her room are another. These two possibilities are physically alike and so—if physicalism is true—alike *simpliciter*. But, as Mary realizes as soon as she leaves her room, this is not the case: the color vision of people outside the room, unlike the color vision of people described in the textbooks, is accompanied by qualia. Since *ex hypothesi* Mary already has all the physical information about people one can have, it follows that the information that distinguishes between these "possibilities" is *not* physical information (Lewis calls it "phenomenal information," but denies the existence of such a thing. We will come back to this in section 4). But this means that physical information leaves open whether two physically alike "possibilities" are alike *simpliciter*, and so physicalism is false. Since the supervenience thesis is, according to Lewis, the bare minimum that is common to all versions of physicalism, and since the KA directly refutes (or, at least, attempts to refute) the supervenience thesis, it follows that the KA goes against the core of physicalism.

The psychophysical conditional

If physicalism is interpreted as a supervenience thesis, it follows that if physicalism is true, the psychophysical conditional is necessarily true. The psychophysical conditional captures the physicalist idea that the psychological features of our world supervene on the physical ones. Returning to the terminology employed in the previous paragraph: since (according to physicalism) any world that is a physical duplicate of the actual world is a duplicate *simpliciter*, it follows that any physical duplicate of our world is also a psychological duplicate. Let P represent the conjunct of all the physical truths of our world and let Q represent the conjunct of all the psychological truths: the supervenience thesis claims that every world at which P is true is a world at which Q is true as well, which means that P entails Q—this is the psychophysical conditional (see Jackson, 1994/2002, p. 165). Hence, if physicalism is true, the psychophysical conditional is necessarily true.

Now, a further question is whether the necessity of the psychophysical conditional must be interpreted as being an a priori or an a posteriori necessity. For the purposes of the KA, it is crucial that it be interpreted as being a priori, because the whole thrust of the argument is to show that the psychophysical conditional is *not* true a priori—that physical truths do not entail certain psychological truths a priori —and infer from this that the conditional is not necessarily true (which amounts to denying supervenience). Of course, the assumption of the aprioricity of the conditional would be completely ad hoc if it wasn't supported by an independent argument. The independent argument is provided by Jackson in "Finding the Mind in the Natural World" (1994/2002). The argument hinges on Jackson's contention that conceptual analysis is indispensable for establishing relations of identity or relations of supervenience between two prima facie independent phenomena (water and H_2O, pain and C-fibers firing, etc.). The role that Jackson assigns to conceptual analysis is not to discover all by itself (purely a priori) which pairs of concepts are identical (he is not claiming that just by thinking about the concept of water we would arrive at the concept of H_2O); rather, he claims that "the very business of conceptual analysis [is] to address which matters framed in terms of one set of terms and concepts are made true by which matters framed in a different set of terms and concepts" (1994/2002, p. 165).

Consider the case of water. Our concept of water is that of a substance that fills the oceans, boils at 212º F at sea level, is necessary for life, is colorless, etc. Empirical inquiry led to the discovery that water

is composed of two atoms of hydrogen and one atom of oxygen, and that this molecule exhibits the behavior that we attribute to water, although perhaps described in different terms (for instance, at 212º F at sea level the substance composed of H_2O changes its physical state; the H_2O molecule sustains life in such-and-such way, etc.). I take it that the role that Jackson assigns to conceptual analysis in the passage just quoted is to "bridge" concepts that are described in different terms (water and H_2O, for example) and determine, guided by what he calls a "principle of charity" (1994/2002, p. 166), whether sentences in which the one appears entail sentences in which the other does. In effect, Jackson thinks that the inference from "Over 60% of the Earth is covered by H_2O" to "Over 60% of the Earth is covered by water" is *not* a posteriori (1994/2002, p. 167]). What is a posteriori is the discovery that H_2O fills the water role, but once one is armed with this piece of empirical knowledge, one can make the above inference *a priori* by conceptual analysis alone.

An important point to notice is that Jackson's conception of the role played by conceptual analysis in making inferences of the type described above is directly relevant to issues of supervenience and reduction. According to Jackson, conceptual analysis allows us "to address the question of whether some inventory of fundamental ingredients does, or does not, have a place for [cases of] *K*s [where *K* denotes an arbitrary property]" (1994/2002, p. 166). So, the role assigned to conceptual analysis is to determine whether the story of the world told in one set of terms has room for certain properties (couched in different terms) we are interested in. It follows—always according to Jackson—that, for example, if physicalism is correct, then there must be a conceptual analysis of the concepts involved in phenomenal truths which render them *a priori* entailed by physical truths. So Jackson is directly denying the possibility of the "explanatory gap" (Levine, 1983) between two domains in which relations of supervenience or reduction allegedly hold: if one cannot see a priori an entailment between sentences couched in a different set of concepts, then one is barred from assuming that the properties described by one of the sets supervene on or are reduced to the properties described by the other. (We will see in section 5 that Loar attacks this conclusion.)

To return to the psychophysical conditional: *if* Jackson's argument about the role played by conceptual analysis in detecting relations of supervenience is correct, then the physicalist (who maintains the supervenience of the mental on the physical) is committed to there being an a priori entailment between physical and psychological concepts (which, as in the case of water, is mediated by the empirical discovery

of the particular physical realizers of specific psychological properties). And *this* is precisely what the KA purports to deny, namely, that physical concepts entail a priori certain kind of psychological concepts, namely phenomenal concepts. If there is no a priori entailment, then it follows (according to Jackson) that phenomenal properties or qualia do not supervene on physical properties, a conclusion that refutes physicalism. Here we appreciate a central feature of the KA: it argues from an epistemic gap (phenomenal concepts are not entailed by physical ones) to a metaphysical gap (phenomenal properties are distinct from physical properties).

Each step of Jackson's argument for the aprioricity of the psychophysical conditional can be called into question, of course. But I think that its main thrust—namely, that an explanatory gap between phenomenal and physical concepts, in case there is one, entails an ontological gap between the psychological and the physical—is extremely persuasive. When we come to discuss Loar's argument in section 5 in which he denies this entailment from the epistemic to the ontological we will appreciate the high costs of doing so and then the persuasiveness of Jackson's case for the aprioricity of the psychophysical conditional will be clearer.

3. Challenging the First Premise of the KA: Stoljar's Two Conceptions of the Physical

A basic tenet of the thought experiment on which the KA is based is that it is possible for Mary to acquire exhaustive knowledge of the physical world from her black and white room, including a complete physical description of color vision in humans. The problem for physicalism arises from the fact that, after being released and seeing color for the first time, Mary realizes that her exhaustive physical knowledge did not contemplate the qualia instantiated by the experience of seeing color, from where one is supposed to conclude that physicalism is false. But if we deny that Mary knew all physical truths while being locked up in her room the argument will not go through: it will remain an open question whether what Mary learns when she sees red for the first time is physical or not.

An intriguing argument for the conclusion that Mary did not have complete physical knowledge before leaving her room is presented by Stoljar (2001/2002). Stoljar's main point is that there are actually two conceptions of what a physical property is: under one of them the KA goes through, but not under the other. These two conceptions

are, respectively, the *theory-based conception* and the *object-based conception*. According to the former, a physical property (a t-physical property) is a property that either appears on physical theory or else supervenes on a property that does appear. According to the latter, a physical property (an o-physical property) is a property which either is the sort of property required by a complete account of "the intrinsic nature of paradigmatic physical objects" or else supervenes on a property required for a complete account (2001/2002, p. 313). Stoljar claims that Mary (before her release) has complete physical knowledge in the first sense (concerning t-properties), but *not* in the second (concerning o-properties). If this is so, the KA would only prove that qualia do not supervene on t-physical properties (because qualia are not entailed a priori by t-physical concepts), but not that they are not physical properties *tout court* (they *may* supervene on o-physical properties and therefore being entailed a priori by o-physical concepts).

What reasons does Stoljar offer for accepting the proposed distinction? He offers two (2001/2002, pp. 313-314]): first, physical theory is concerned only with the *dispositional* properties of physical objects and is silent about their *categorical* properties. (A good example, which Stoljar takes from Blackburn, is mass: at first, mass may seem the clearest example of a categorical property, but actually it is knowable only through its dynamical effects.) Second, dispositional properties require categorical grounds in order to be instantiated. (Think of a fragile vase: the property of being fragile is dispositional, but there must be a non-dispositional property or properties in virtue of which the vase is fragile.) *If* one accepts both reasons, a straightforward argument for the distinction between t-properties and o-properties follows (Stoljar, 2001/2002, pp. 320-321): if you are a physicalist who believes only in t-properties (that is, if the only physical properties you recognize are the properties postulated by physical theory) *and* at the same time accept the metaphysical thesis that dispositions require categorical grounds, then you are committed to the idea (which amounts to a negation of physicalism) that each time a physical (dispositional) property is instantiated, a non-physical property is instantiated as well, because the categorical grounds of dispositional properties are *not* part of physical theory and so, from the point of view of t-physicalism, are not physical properties.

So we have an independent argument for the distinction between t-physical and o-physical properties, a distinction that undercuts the KA right in the first premise: Mary (before her release) does *not* know everything physical there is to know about other people, because she

only has complete knowledge about t-properties, not about o-properties (Mary's knowledge comes from physical theory, which speaks only about t-properties). Since o-properties *are* physical properties, it follows that Mary does not have complete physical knowledge *tout court*. But then it is at least possible that what she learns when she sees red for the first time is something physical (or something that supervenes on a physical o-property). Hence, if physicalism is interpreted as encompassing both t- and o-properties, the KA does nothing to show that physicalism is false.

Let us concede for the sake of the argument that Stoljar is correct in distinguishing between t- and o-properties and that the distinction falsifies the first premise of the KA. Now the question is: what are exactly o-properties? Are they qualia? There is an apparently obvious route to the conclusion that qualia are at least *one* sort of categorical or intrinsic property, namely, that the concept of a categorical property seems to be modeled on the concept of qualia (Stoljar, 2001/2002, p. 321). Take an example: the "what it is like" of the experience of seeing red seems to be intrinsic to this experience, in the sense that it cannot be accounted for in dispositional terms: while it is true that when I see red I am disposed, for instance, to utter the words "I see red" if questioned about what color I am seeing, the verbal report neither captures the phenomenal experience I am undergoing nor is identical with it. The hypothesis of the inverted spectrum is useful in this respect: you and I may utter the words "I see red" when confronted with a ripe tomato, even though it is possible that the quale you name "red" corresponds to the quale I name "green." So, it seems that the phenomenal aspect of our color experience cannot be captured in dispositional terms because our verbal dispositions are identical while the intrinsic character of our experiences is (or may) not.

However, Stoljar himself dismisses the force of this argument on the grounds that it does not follow, from the fact that our concept of a categorical property is modeled in our concept of qualia (something that can be disputed), that categorical properties are qualitative properties themselves. (If it did follow, we would be committed to panpsychism, a stance that Stoljar roundly rejects.) But if o-properties are not qualia, what are they? Stoljar's definition is not very informative: "these are [the physical and non-qualitative] properties which make up the categorical nature of physical objects" (2001/2002, p. 321).

But the problem confronted by Stoljar is not just circularity, but something worse: o-properties are (at least in our current scientific stage) ineffable. Stoljar explicitly recognizes this: "o-physicalism is

[committed] itself to a class of truths which cannot be expressed in a language we currently understand" (2001/2002, p. 321). The problem is straightforward: we pick out t-properties by using t-concepts, phenomenal properties by using phenomenal concepts, and o-properties by using… what? Well, o-concepts. But what are these? We do not know, because o-concepts are neither the concepts employed by (current) physical science nor the concepts we employ to refer to our phenomenal experiences. O-concepts are a sort of postulate that the physicalist makes in order to remain true to the idea that if physicalism is true, it is a priori true: "What our position predicts is that in order to have an a priori physicalist theory of qualia and their place in the world … one would need to complete the categorical inquiry" (Stoljar, 2001/2002, p. 322), an inquiry that, presumably, would employ o-concepts.

It may seem unfair to call o-concepts a postulate, given the independent metaphysical reason Stoljar gives to believe in o-properties (the disposition/ground distinction). After all, if we have reason to believe that o-properties exist, it seems reasonable to say that we have a reason to believe in o-concepts. True enough. The problem is that we do not have the slightest idea of what these concepts would be; even worse, as Stoljar admits, "the distinction between categorical and dispositional [properties] does not seem to matter much to the main business of science" (2001/2002, p. 321), so it is possible that we will *never* know what o-concepts—and *a fortiori* o-properties—are. (Unless, of course, one claims that a discipline apart from science would be in charge of investigating o-properties. But what would that discipline be?)

Stoljar has a good point in that, if we accept the metaphysical distinction between disposition and ground, and consequently come to believe in o-properties as being something different from t-properties, the KA becomes powerless against physicalism. Maybe Mary only has t-knowledge, which is not all the physical knowledge there is to have; but the problem is that the rest of us are in the same position as Mary, and for all we know so will be everyone in the future. So, while Stoljar's two concepts of the physical prevent the KA from refuting physicalism, it does so at the cost of postulating a realm of semi-noumenical properties beyond the reach (or interest) of science, while insisting that, for all we know, these could (or must?) be physical properties. Hence, the main service Stoljar's makes to physicalism is to insulate it from the KA by leaving open the possibility that the knowledge that Mary acquires when she leaves the room could be inferred a priori from the mysterious o-concepts. However, it seems that, by pursuing this strategy, Stoljar has not shifted the burden of proof to his opponent, since he still owes

us a much more substantive account of o-concepts.[2] Until we have such an account, I think that we should explore other options for resisting the KA.

4. Challenging the Second Premise of the KA: Lewis' Ability Hypothesis

A different strategy against the KA is to attack its second premise, namely, that Mary learns something about other people and herself after being released—something that was not captured in her comprehensive knowledge of physics. Philosophers who opt for this line of attack usually do not deny outright that Mary learns *something*, rather they deny that *what* she comes to learn constitutes a kind of knowledge that threatens physicalism and the supervenience thesis.[3] This is the strategy taken by Lewis in "What Experience Teaches" (1988/2002). The main thesis of the paper is that the knowledge that Mary gains upon seeing colors for the first time is knowledge-how, not knowledge-that (1988/2002, p. 293). Specifically, Lewis puts forward the bold idea that "knowing what an experience is like *just is* the possession of these abilities to remember, imagine, and recognize. It isn't the possession of any kind of information, ordinary or peculiar" (1988/2002, p. 293). Lewis' idea is this: before Mary saw red for the first time, she was unable to remember, imagine or recognize red, but once she saw red, she immediately acquired these abilities. And that is all that Mary learned, according to Lewis, when she left the room: she gained know-how about remembering, imagining, and recognizing instances of red.

Of course, paraphrased in this way, the Ability Hypothesis seems like an instance of the strategy of "sticking your head in the sand," in this case, refusing to acknowledge qualia. The qualia freak can agree with Lewis that Mary acquires all the mentioned abilities when she sees red for the first time, but she would insist that Mary acquires them *by* getting in touch with a *sui generis* kind of information, namely, phenomenal information.[4] To block this rebuttal, Lewis offers

[2] To clarify, my point is not that Stoljar has had nothing to say in defense of o-properties, but rather that, since by his own admission these are (currently at least) beyond the reach of physical science, appealing to them isn't the best way to challenge the KA—specifically because they do nothing to shift the burden of proof to the antiphysicalist.

[3] Dennett (1991/2004), for example, does deny that Mary learns anything.

[4] This is precisely the rebuttal offered by Martine Nida-Rumelim (1995/2004, p. 258). She speaks of phenomenal knowledge instead of phenomenal information. I

a thorough criticism of the very notion of "phenomenal information," showing that, despite its initial plausibility—i.e., the representation of that aspect of experience which cannot be conveyed through discursive lessons—the notion is deeply problematic. I will argue that discrediting the notion of phenomenal information is the best service Lewis does to physicalism in the aforementioned paper, because his Ability Hypothesis is unconvincing for reasons to be discussed below.

Lewis' strategy to discredit the KA is to show that it assumes in its second premise that the Hypothesis of Phenomenal Information (HPI) is true. Only if we grant this hidden assumption and concede that what Mary acquires is phenomenal information, we can infer the conclusion of the KA—that there are truths (information) about other people that escape the physicalist story. But why grant the HPI? The hypothesis presents itself as the obvious candidate for explaining why Mary could not know in her black and white room (despite having exhaustive knowledge of physics) what she came to learn once she was released. In effect, the HPI is the main leverage behind the intuition-pumping mechanism trying to convince us that the physicalist story cannot capture qualia. Take an example: Mary knows all about ripe tomatoes, including the wavelength they reflect and the way the visual system of human beings reacts when confronted with this wavelength. So Mary had complete physical information about the visual relations between humans and ripe tomatoes. Still, the qualia freak claims, she missed something, namely, the phenomenal information that human beings acquire when visually confronted with a ripe tomato; this is the information she acquires when she leaves the room and has the experience of seeing a ripe tomato. Thus, the qualia freak concludes, this must be a *sui generis* kind of information since it cannot be conveyed by physical knowledge alone.

The point of calling "information" what Mary acquires after being released is not gratuitous. As Lewis notes, this is what allows the qualia freak to mount her case against physicalism, because genuine information allows us to "eliminate possibilities" (1988/2002, PP. 287-288]). Suppose you are travelling by subway, fall asleep and minutes later wake up without knowing which station you are about to arrive at. You quickly consider various possibilities after guessing how much time you slept; these are all live possibilities until it is announced that the train is arriving to station X. Armed with this information, you proceed to discard options Y and Z. According to the qualia freak,

expand on Nida-Rumelin's position in fn. 8 below.

something similar occurs to Mary: as it was discussed in section 2, we can think of the world outside her room and the world as depicted in her physics books as two different possibilities. Before leaving the room, Mary thinks of both these possibilities as physically identical, which means that no amount of physical information would discriminate one but not the other (this is the supervenience thesis). However, when she sees color, she realizes that the actual world is not the world described by physical information, since this information left something out. But, the qualia freak argues, the only way for Mary to discriminate between these two possibilities (the "physics world" and the "outside-the-room world") is by way of acquiring information, which ex hypothesi cannot be physical information. Hence, it must be a different kind of information, a kind that allows Mary to discriminate between two physically identical possibilities—a feat which directly contradicts the supervenience thesis and physicalism with it.

But what is wrong with phenomenal information? Lewis presents two main reasons for rejecting the HPI. The first appeals to the fact that, were the hypotheses true, it would prove too much, because it would show that phenomenal information is beyond not only the physicalist story but beyond *any* story stated in propositional terms. To illustrate the point Lewis introduces "parapsychology," an imagined science encompassing the whole of non-physical entities, properties and processes. Can parapsychology capture phenomenal information? Obviously not,

Lewis argues, because a parallel KA can be run against parapsychology: were Mary to become a leading expert in the field of parapsychology, she still would not know what it is like to see red before *experiencing* the sensation of red. So phenomenal information is not to be equated even with parapsychological information, but it is supposed to be information, nonetheless. But what kind of information is it? As in the case of Stoljar's o-properties, phenomenal information seems to belong to the realm of the ineffable. Lewis claims that if we find an alternative interpretation of the kind of information/knowledge imparted by experience that *can* be characterized in positive terms, we should prefer the alternative. But the latter task does *not* belong to the physicalist: once she has criticized the commonsensical notion of phenomenal information, she shifts the burden of proof to the antiphysicalist, who is now charged with the task of explaining what kind of information Mary acquires, or so I will argue now.[5]

[5] An anonymous referee has worried that the antiphysicalist isn't going to be

How much force does Lewis' argument against the HPI have? In "What Mary Didn't Know" Jackson thinks that not much (1986/2004, p. 55). There he argues against an identical criticism earlier posed by Churchland and claims that there is no "parity of reasons" between physicalism and parapsychology (or dualism as Churchland calls it) concerning the conclusion of the KA, because the first premise of the argument turns out to be false if we substitute "parapsychology" for "physicalism" so as to read: "Mary (before her release) knows everything parapsychological there is to know about other people." The premise is false, according to Jackson, because it is not plausible that a complete parapsychological story could be learned inside a black and white room. This response is obviously inadequate, because it just *insists* that a complete parapsychological (or dualist) story includes qualia, while failing to provide a substantive characterization of that story that goes beyond this fact. Compare: the physicalist lays her cards on the table by clearly stating what she means by physical information (information included in physical theory); by contrast, the qualia freak just insists that phenomenal information is whatever Mary acquires when she leaves the room and denies that it can be explained in any other way—not even a parapsychological one. But that amounts to jumping out of the frying pan into the fire: in order to protect the HPI from a parallel version of the KA, the qualia freak bites Lewis' bullet and tacitly accepts that phenomenal information is ineffable—it cannot be explained but you just know when you have it. By doing so, the qualia freak faces the charge of being deliberately obscure in order to save her position.

The second reason for rejecting phenomenal information that Lewis provides is that, in case there was any, it would necessarily be epiphenomenal. Suppose we try to deny this and claim instead that phenomenal information causes certain changes in the physical world, for example, that it causes Mary to say, "Now I can see I was wrong in thinking I knew everything." What is the problem with this? Well, this would amount to a rejection of the causal closure of the physical world, because a non-physical entity (qualia) would cause a physical change (Mary's utterance). Given the successes of physics, the qualia freak needs to offer a pretty powerful and independent argument for rejecting the causal closure, an argument that goes beyond the mere insistence

impressed by my shifting the burden of proof to them concerning the explanation of the notion of phenomenal information. I completely agree; however, my goal in this paper isn't to convince the antiphysicalist to surrender her position, but rather to provide tools for an "impartial arbiter" to adjudicate the dispute or, at the very least, to determine which of the two parties has the advantage in it.

that phenomenal information exists. Now suppose the qualia freak bites the bullet again and rests content with the fact that phenomenal information is epiphenomenal (as Jackson did in the original 1982 article). This only makes her position more untenable, because now it is utterly unclear how she manages to even talk (let alone write articles) about phenomenal information. Presumably, talking and writing about qualia implies that one has some traces of them in one's memory; however, leaving a trace in one's memory is a physical change, one that cannot be brought about by something lacking causal powers.[6]

As I said above, I think that Lewis has a convincing argument against the HPI; an argument that, moreover, shifts the burden of proof to the qualia freak: she must now provide a positive characterization of what exactly she is talking about when she talks about "the phenomenal aspect of experience." I will come back to this point in sections 5 and 6. In the meantime I want to offer a criticism of Lewis' own answer to the question of what Mary learned when she left her room—the Ability Hypothesis (AH).

Recall that Lewis claims that what Mary acquires when sees color for the first time is knowledge-how—a set of abilities to recognize, remember and imagine certain experiences—and not knowledge-that––a set of propositions concerning a special, non-physical side of reality. Since abilities cannot be acquired just by taking discursive lessons, this explains why Mary learned something when she was confronted with red for the first time, something that she could not have learned just by reading about the physical properties of color vision. Would the qualia freak accept Lewis' characterization of Mary's new knowledge? Of course not, and I think she would have good reasons to offer Lewis because whatever changes occurred in Mary when she saw red, they were beyond the mere acquisition of certain abilities.

Let me begin by criticizing an attack to the AH that does *not* work. Loar (1997/2002, p. 304) claims that the AH cannot be true and that we should accept the idea of phenomenal concepts instead because we can produce sentences in which thoughts about the phenomenal aspect of experience interact with straightforward propositional content. If this is so, and the resulting sentences are meaningful, then phenomenal thoughts must have predicative content, so they are not just reports on the execution of certain abilities. For example, consider

[6] It is noteworthy that, by his own admission, this was the argument that convinced Jackson that his characterization of qualia was wrong and led him to embrace physicalism. See "Postscript on Qualia" (1998/2004).

the sentence: "If apples taste like *this*, then my mother was right." Loar argues that sentences like this are fully meaningful; moreover, this one has the arrangement of an inference, so the antecedent must have some sort of content. Does this argument show that the antecedent in sentences like this one cannot be occupied by reports of abilities? I doubt it. As a counterexample, consider the following sentence, which has embedded in it an ineliminable report of ability: "I can make *this* movement, therefore I am not paralyzed." This sentence is, I think, as meaningful as the former. Loar might want to reply that in this sentence the report can be replaced without remainder by a description like, for instance, "I can move my arm in such-and-such a way". I do not think this would be right, though, because descriptions of bodily movement are insufficiently fine-grained to capture ability reports *in just the same way* that descriptions of phenomenal experiences are insufficiently fine-grained to capture reports of undergoing particular phenomenal experiences (*e.g.,* "If apples taste like *such-and-such*, then my mother was right"). Since Loar's argument is that, since the latter sort of reports are ineliminable and yet the resulting sentences are meaningful, phenomenal thoughts must have predicative content; by parity of reasoning I conclude that (at least some) reports of abilities are ineliminable and yet the sentences in which they are embedded are meaningful because such reports also have predicative content. Therefore, in a sentence like "I can make *this* movement, therefore I am not paralyzed", we can take the report at face value as making direct reference to the ability of performing *this* movement (which the agent instantiates as he talks). So, my claim is that reports of abilities *can* interact with propositional content to form meaningful sentences.

But then what is wrong with the AH? To begin with, the qualia freak can complain that Lewis' statement of the AH ("knowing what an experience is like *just is* the possession of these abilities to remember, imagine, and recognize") is misleading because the abilities mentioned are abilities to remember *what it is like* (to see red), to imagine *what it is like* (to taste Vegemite), etc. Lewis apparently reduces experiential knowledge to these abilities, but the reduction seems to go through only because it fails to mention what these abilities are about. And what they are about, the qualia freak would say, is exactly what she is trying to capture with notions such as "qualia", "raw feel", etc. These may be obscure notions, but they certainly point to something different from the acquisition of abilities. This response is sort of question-begging, especially because, as we have seen, the qualia freak lacks a satisfactory characterization of phenomenal information; still, I think that there is

something right in her complaint. We can ask Lewis: Mary acquires the ability to remember… what?

Lewis' answer is that these abilities are related to experiences: Mary, after her release, can remember the experience of red, imagine her experience of red, recognize other instances of the experience of red, etc. But, the qualia freak would say, these experiences are *not* just the abilities thus acquired. Consider the following analogy: after learning the concept "two," I acquired certain abilities like recognizing that the number of apples in the table is two, imagining two chairs, remembering having seen two blue cars, etc. Are we tempted to say that the concept "two" is nothing but these abilities? Obviously not. But then, why should we accept that experiences are nothing but the abilities Lewis mentions? (Lewis could respond that there is no analogy here, since we have at least an idea about how to understand concepts about numbers, whereas we lack any clear conception of the phenomenal aspect of experience.)

A couple of stronger objections to the AH are the following: first, it is conceivable that a person might have some sort of intellectual deficiency impairing her to acquire the abilities to recognize, remember, and imagine certain experiences; if such a case is possible (think, perhaps, of patients with Alzheimer disease or persons who have lost short term memory), are we going to say that the person cannot know what any experience is like? I do not think this would be the correct description of the case. Suppose we give to such a person Vegemite for lunch. Her sensory apparatus is still working properly, and we might even observe some behavioral evidence that she has perceived the taste (perhaps she made a grimace of disgust); should we say that she does not know what Vegemite taste like because she has not acquired the abilities listed by Lewis? I feel strongly inclined to answer in the negative. Could Lewis come back and say that the abilities are there, but just cannot be exercised? This would not work: a person who becomes blind simply loses his ability to see and, similarly, the person we are imagining has lost her capacity to deploy the mental abilities in question; still, it seems that she *can* know what an experience is like. Hence, *contra* Lewis, knowing what an experience is like is not just the possession of the abilities to remember it, imagine it, and recognize another instance of it.

Second, the AH fails to do justice to a central feature of Jackson's version of the KA: Mary learns something about *other* people, not just about herself, namely, that *they* experience the world in a way she was ignorant of. Can this feature of the KA be adequately captured

by the AH? I think not. For consider: is the knowledge Mary acquires knowledge about a set of abilities that other people have and which she knew nothing about? Clearly not; after all, Mary studied human psychology from her room, so she already knew that people have the abilities to remember, imagine and recognize their experiences. What she ignored, according to the KA, is a certain way in which the subject matter of these abilities presents itself. Again, Lewis might complain that the qualia freak lacks any positive description of this special mode of presentation, but what is clear from these two last objections is that, whatever it might be, it resists being reduced to mere abilities.[7]

5. Challenging the Relevance of the KA's Conclusion: Loar's Phenomenal Concepts

As I explained in section 2, the KA attacks physicalism by moving from an explanatory (or epistemic) gap to a metaphysical one. In doing so, the argument assumes that the supervenience of the mental on the physical must hold a priori; if it does not, then there is no such supervenience. The two responses we have explored thus far (Stoljar's and Lewis') accept the aprioricity of the psychophysical conditional and consequently attempt to deny that an explanatory gap opens between phenomenal and physical concepts. The third and last response we will examine bites the antiphysicalist's bullet and accepts the explanatory gap but denies that we can infer from it a metaphysical one. A clear example of this sort of response is found in Loar's paper "Phenomenal States" (1997/2002). Loar's main insight is that phenomenal concepts constitute a separate class distinct from, and irreducible to, physical

[7] An anonymous referee argued that this objection is question-begging because Mary cannot simply assume that other people undergo a similar kind of experience—i.e., a phenomenal experience—as herself upon seeing red. In response, notice that the assumption that Mary learned something about the experiences of other people presumes only that Mary is neurotypical and therefore she can make the defeasible inference that other people (probably) undergo phenomenal experiences like herself (even though she cannot be sure that they are the same in phenomenal character). This inference is compatible with the point made by the referee to the effect that, according to the qualia freak, qualia are eminently introspective, since it relies only on the possibility of other people having phenomenal experiences of one sort or another, although not exactly the same as she is undergoing (the said inference is compatible, for instance, with the inverted spectrum hypothesis). If the KA were a skeptical argument against the existence of other minds, I would agree that the inference would be question-begging, but the KA isn't usually read this way and so it seems to me that we are allowed to extend the argument in the direction suggested in the text.

concepts. The former are type-demonstrative concepts "that derive their reference from a first-person perspective" (1997/2002, p. 295), while the latter are third-person concepts that pick out (in the case of mental properties) functional states. The crucial point is that, according to Loar, the referent of phenomenal concepts is the *same* as the referent of physical ones, namely certain brain states. What differs between the two kinds of concepts is the *mode of presentation* of its referent, in much the same way that "Superman" and "Clark Kent" pick out the same referent under different modes of presentation: "the properties these [phenomenal concepts] *phenomenologically reveal* are physical-functional properties—but not of course under physical-functional descriptions" (Loar, 1997/2002, p. 299). Since phenomenal concepts are irreducible to physical ones, the identity of phenomenal and physical properties is, according to Loar, necessary but a posteriori (contrary to Jackson's argument presented in section 2). The a-posteriority of the identification explains, then, the unbridgeable explanatory gap between phenomenal and physical concepts, while its necessity accounts for the absence of a corresponding metaphysical gap.

Now, the obvious difference between the cases of "Superman" and "Clark Kent" on the one hand and, say, "pain" and "C-fibers firing" on the other is that no one would claim that an explanatory gap opens in the former, while many antiphysicalists (and some physicalists like Loar) claim that such gaps exists in the latter. Loar explains the asymmetry by arguing that phenomenal concepts belong to a class he labels "recognitional concepts" which resist assimilation to any other class. Recognitional concepts have the form "X is one of *that* kind" (Loar, 1997/2002, p. 298). Since the reference to the encompassing kind is demonstrative, Loar claims that recognitional concepts are type-demonstratives, "grounded in dispositions to classify, by way of perceptual discriminations, certain objects, events, situations". The main reason Loar offers to explain why this class of concepts is irreducible to any other is that, as mentioned in the previous paragraph, in the case of recognitional concepts the referent is fixed from the first-person perspective. In other words, it is only from the first-person perspective that the classification of particulars as belonging to certain kinds is made. As it happens, in the case of phenomenal concepts the relevant encompassing kind is in each and every case a physical-functional property, even though the person doing the classification does not need to know it is so. For instance, when I say: "this color sensation belongs to *that* kind" while seeing a patch of red, I do not think of the relevant kind as such-and such brain state (rather, the kind I am thinking about

is also phenomenal), even though, unbeknown to me, it *is* such-and-such brain state.

That two completely different and irreducible kinds of concepts refer to the same properties already sounds like an act of serendipity, but what interests me for present purposes is that Loar's argument shows how difficult it is to make sense of the relevant class of concepts that, allegedly, characterize phenomenal experience. Although Loar is a physicalist, he delineates more clearly than most antiphysicalists what a phenomenal concept would be.[8] A phenomenal concept has two main characteristics: 1) it picks out essential properties of its referent not by way of contingent modes of presentation; 2) it is essentially a recognitional concept, the referent of which is fixed from the first-person perspective. The first point is meant to capture the intuition that, since seeing a patch of red *as* a patch of red is an essential feature of this experience, the concept that refers to this feature is capturing something essential to it. The second point appeals to the fact that the only way of acquiring phenomenal concepts is by way of being confronted with the relevant experiences.

Notwithstanding Loar's insistence to the contrary, the first point seems plainly inconsistent with physicalism because it is utterly mysterious how one and the same property (C-fibers firing, say) can be captured by way of a phenomenal (pain) concept ("X is a sensation of *that* kind") and by a physical-functional concept, in spite of the fact that both kinds of concepts are irreducible to one another and that *both* pick out an essential property of its referent (not by way of a contingent mode of presentation). According to Loar's picture, we just must accept that "feeling like *this*" and "C-fibers firing" pick out the *same* functional

[8] A similar proposal but by a philosopher with antiphysicalist sympathies is Martine Nida-Rumelin's (1995/2004), although she speaks not of phenomenal concepts but of phenomenal belief and phenomenal knowledge. Her proposal is "to attach the subscripts 'p' for ('phenomenal') and 'np' (for 'non-phenomenal') to color terms within belief contexts to express the intended distinction" (1995/2004, p. 245). Thus, for example, Mary (before her release) believes$_{np}$ that the sky is blue, while (after her release) she believes$_{p}$ that the sky is blue. Nida-Rumelin claims that the latter belief can only be acquired after having undergone the relevant experience, and hence that it is a kind of belief that cannot be obtained through discursive lessons. (In the paper quoted she does not take the further step of claiming that phenomenal belief/knowledge entails the existence of non-physical facts. Hence, her strategy is compatible with Loar's, that is, accepting an epistemic gap while denying or remaining agnostic about a metaphysical one.) I think that the notion of phenomenal belief/knowledge suffers from the same problems (discussed below) that afflict Loar's phenomenal concepts.

property, despite the fact that we cannot find a connection between these two essential modes of presentation of the same functional property.

But my main quarrel with Loar's phenomenal concepts has to do with the second point—with the way this kind of concepts is supposed to pick out its referent. Loar claims that the referent of a phenomenal concept is fixed from the first-person perspective, that is, the referent is fixed when the person employing the concept gets acquainted with the relevant kind. The example he offers is the following: you can acquire the concept "porcelain" by way of a description of the relevant features of the kind porcelain and then learn to recognize instances of this kind. Is this a recognitional concept? Loar says it is not; a recognitional concept must be "recognitional at its core; the original concept is recognitional" (1997/2002, p. 298). I take him to mean the following: when you acquire the third-person concept of porcelain, the referent is fixed by a relevant description; but then you may acquire a *different* concept of the form "X is of *that* kind," the referent of which being fixed when you get directly acquainted with the X in question (in this case, a sample of porcelain).

Now let us apply Loar's reasoning to the case of phenomenal concepts. It would go like this: Mary has, before her release, one concept of "red" (couched in the language of physics) and then, after she sees red for the first time, she acquires a second, recognitional concept of "red" (she will think "X [a patch of red] is one of *that* kind [presumably, the kind 'qualia' or, alternatively, 'red qualia']"), the referent of which is fixed by her own visual experience. If this is the correct way of understanding Loar's suggestion, then I think that the very idea of phenomenal concepts (as well as the idea of phenomenal belief, see fn. 8 above) is unappealing for the following two reasons:

First, it implies that we have *two* different concepts (or beliefs) for *every* property that can be perceived through the senses, one corresponding to its propositional description and another one corresponding to the "recognitional" acquaintance with it. If this were the case, it would be an amazing coincidence that the public and the private versions of these concepts (or beliefs) cohere so well in ordinary life as they apparently do. I think it is much more plausible to account for the correspondence among the reports of different people concerning sensible qualities by positing just one publicly accessible concept for each quality.[9]

[9] An anonymous referee has objected that by "recognitional concept" Loar doesn't mean a private concept "in the sense that its uses would be beyond possible correction by others (or by the same subject at different times)". In response, I think that Loar's insistence that such concepts are "recognitional at their core", together with the idea

Second, the very idea of the referent of a concept fixed individually by each person brings with it Wittgenstenian worries concerning rule-following, the impossibility of private language, and the like. Loar acknowledges, and then simply dismisses, these worries, claiming that phenomenal concepts are part of "unanalyzed common sense concerning a natural group of concepts and apparent conceptual abilities" (1997/2002, p. 298) and that problems about the irreducibility of phenomenal properties to physical ones arise from this (commonsensical) perspective. The latter assertion is false since, as we saw in part 2, Jackson offers a *philosophical*, not a commonsensical argument for the a priori entailment of concepts involved in relations of supervenience, and for the conclusion that, when the entailment does not occur, we have reason to believe that we have two distinct properties. Moreover, the problem with phenomenal concepts is *precisely* their commonsensical appearance. The intuition that drives the KA is, as we saw in section 4, that there is a special kind of information (in Loar's terms, a special kind of concepts) that one obtains only through direct experience, a kind that cannot be couched in physical terms. But, when we press the qualia freak to explain what kind of information phenomenal information is supposed to be, or when we try to make sense of phenomenal concepts, we discover how deeply problematic both notions are.

In sum, Loar's strategy for defeating the KA—showing that truths couched in irreducible non-physical vocabulary are compatible with physicalism because they pick out physical properties—is unsatisfactory for two main reasons: on the one hand, it demands that we just accept a brute and unexplainable identity between phenomenal and physical properties; on the other, it relies on the notion of phenomenal concepts, which, as we saw, is deeply problematic in itself. For present purposes, the main interest of Loar's paper is that it illustrates the obscurities and perplexities that accost philosophers as soon as they try to define exactly what Mary learned when she left her room.

6. Conclusion

We have reviewed three different responses to the KA. I have argued that the strongest one consists in attacking premise two—not by denying that *something* happens to Mary when she leaves the room

that for *every* perceptible object we have two different concepts, do invite Wittgenstein worries of the kind discussed in the next paragraph and, moreover, result in an unacceptable proliferation of concepts.

but by shifting the burden of proof to the qualia freak, who must now explain clearly and convincingly *what* it is that Mary learns. We have seen that Lewis' criticism of the notion of phenomenal information casts deep doubts on the intelligibility of that notion; we also saw that Loar's phenomenal concepts, which allegedly would clarify the kind of knowledge acquired by Mary, introduce great difficulties of their own. The core intuition behind the KA is that Mary acquires a piece of knowledge which was inaccessible to her while she was captive in the black and white room. But what exactly is this piece of knowledge about, and why is it inaccessible from Mary's physical conceptual network? These are the key questions the antiphysicalist must answer—and she cannot rest content with a vague appeal to either phenomenal information or phenomenal concepts (or beliefs), at least not in the form discussed in this paper.

The other two strategies we explored for responding to the KA are, I think, much less satisfactory. Stoljar denies that Mary has complete physical knowledge (the first premise of the KA), appealing to the fact that physical theory is concerned only with dispositional properties while leaving intrinsic properties untouched. Hence, according to Stoljar, it is entirely possible that the knowledge Mary acquires is physical knowledge corresponding to the intrinsic properties of matter. The two main problems I found with this strategy are, first, that it postulates a quasi-noumenical realm of intrinsic properties that seems to be beyond scientific inquiry and, second, that it is too close to what Chalmers (2002, p. 265) has called "panprotopsychism"—which is not clearly a form of physicalism. On the other hand, Loar challenges the conclusion of the KA by arguing that the existence of truths that cannot be couched in physical terms is not a problem for physicalism if the concepts that express those truths have physical properties as referents. I also found two main problems with this response: first, that by accepting an unbridgeable explanatory gap between phenomenal and physical concepts, he concedes too much to the antiphysicalist; second, that the very notion of phenomenal concepts is not coherent enough for capturing what it is that Mary learns.

As I said, I am not inclined to deny that Mary undergoes an important change when she leaves the room, but the arguments of sections 4 and 5 make me think that there are severe problems in characterizing this change as the acquisition of a special, non-physical kind of information, concepts or knowledge. Moreover, I have claimed that the task of offering a precise characterization of Mary's presumptive new knowledge corresponds to the antiphysicalist. After

all, the physicalist has a pretty clear notion of the kind of information she regards as knowledge—it is the information of physical theory. By contrast, the antiphysicalist just relies on the (problematic) intuition that the knowledge Mary acquires is non-physical. So, it is not the physicalist who must advance a precise characterization of the situation Mary finds herself in when she leaves the room (for instance, I claimed the Lewis' Ability Hypothesis does not work either). I am very much in agreement with Paul Churchland when he says that the antiphysicalist, in stating her challenge to physicalism, *makes* the problem of irreducible qualia "transcendentally hard at the outset by presumptive and question-begging *fiat*" (1996 [2002: 365]). If my analysis in this paper is correct, the question-begging fiat is the vague notion of phenomenal information or, alternatively, of phenomenal concepts.

References

Chalmers, D. (2002). Consciousness and its place in nature. In D. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 247-272). Oxford University Press.

Churchland, P. (1996/2002). The rediscovery of light. In D. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 362-371). Oxford University Press. (Original work published 1996, *Journal of Philosophy, 93*, 211-228.)

Dennett, D. (1991/2004). 'Epiphenomenal' qualia? In P. Ludlow, Y. Nagasawa & D. Stoljar (Eds.), *There's something about Mary* (pp. 59-68). The MIT Press. (Original work published 1991, D. Dennett *Consciousness explained*. Brown.)

Jackson, F. (1982/2004). Epiphenomenal qualia. In P. Ludlow, Y. Nagasawa & D. Stoljar (Eds.), *There's something about Mary* (pp. 39-50). The MIT Press. (Original work published 1982, *Philosophical Quarterly, 32*, 127-136.)

Jackson, F. (1986/2004). What Mary didn't know. In P. Ludlow, Y. Nagasawa & D. Stoljar (Eds.), *There's something about Mary* (pp. 51-56). The MIT Press. (Original work published 1986, *Journal of Philosophy, 83*, 291-295).

Jackson, F. (1994/2002). Finding the mind in the natural world. In D. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 162-169). Oxford University Press. (Original work published 1994, R. Casati et al. Eds., *Philosophy and the cognitive sciences.* Holder-Pichler-Tempsky.)

Jackson, F. (1998/2004). Postscript on qualia. In P. Ludlow, Y. Naga-

sawa, & D. Stoljar (Eds.), *There's something about Mary* (pp. 417-420). The MIT Press. (Original work published 2004, F. Jackson, *Mind, method and conditionals*. Rutledge.)

Jackson, F. (2004). Mind and illusion. In P. Ludlow, Y. Nagasawa & D. Stoljar (Eds.), *There's something about Mary* (pp.421- 442). The MIT Press.

Levine, J. (1983). Materialism and qualia. The explanatory gap. *Pacific Philosophical Quarterly, 64*, 354-361. https://doi.org/10.1111/j.1468-0114.1983.tb00207.x

Lewis, D. (1988/2002). What experience teaches. In D. Chalmers (Ed.) *Philosophy of mind: Classical and contemporary readings* (pp. 281-294). Oxford University Press (Original work published 1988, *Proceedings of the Russellian Society*.)

Loar, B. (1997/2002). Phenomenal states. In D. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 295-311). Oxford University Press (Original work published 1997, N. Block et al. Eds., *The nature of consciousness.* The MIT Press.)

Nida-Rumelin, M. (1995/2004). What Mary couldn't know: Belief about phenomenal states. In P. Ludlow, Y. Nagasawa & D. Stoljar (Eds.), *There's something about Mary* (pp. 241-267). The MIT Press. (Original work published 1995, T. Metzinger Ed., *Conscious experience,* pp. 219-245. Imprint Academic.)

Pettit, P. (1993). A definition of physicalism. *Analysis, 53*, 213-223.

Stoljar, D. (2001/2002). Two conceptions of the physical. In D. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 311-327). Oxford University Press. (Original work published 2001, *Philosophy and Phenomenological Research, 62*, 253-281.)

Received 16th September 2021; revised 31st January 2022; accepted 25th March 2022.