2012

Mary's Dilemma: A Novel Take On Jackson's Famous Thought Experiment

Noah O. Abolafia-Rosenzweig
Claremont McKenna College

CLAREMONT McKENNA COLLEGE

MARY'S DILEMMA:
A NOVEL TAKE ON JACKSON'S FAMOUS THOUGHT EXPERIMENT

SUBMITTED TO

PROFESSOR AMY KIND

AND

DEAN GREGORY HESS

BY

NOAH ABOLAFIA-ROSENZWEIG

FOR

SENIOR THESIS

FALL 2012
3 DECEMBER 2012

1. Introduction

Physicalism, defined as the belief that everything which exists can be accounted for by a completed physics, faces one of its most serious challenges in a thought experiment involving a color scientist named Mary. Famously devised by Frank Jackson in "Epiphenomenal Qualia," the thought experiment intends to deny that certain mental phenomena are physical. Specifically, when he created the experiment, Jackson wanted to claim that things like the painfulness of pain and the itchiness of itches—the way experiences are like for the experiencer—what Jackson refers to as *qualia*—cannot be accounted for science. Though he has since recanted his original position, Jackson's famous thought experiment retains its influence and is a primary hurdle for any physicalist.

As a result of its popularity, many physicalists have attempted to craft positions which avoid the conclusion of Jackson's thought experiment. In section 2, I will examine three such attempts which represent a large body of work on the subject. Each, I will argue, contains a fatal flaw. I will note the faults of the physicalist theories discussed, and, in section 3, I will develop a new position informed by their errors. In the end, I hope to show that, while many attempts at shirking Jackson's conclusion have failed, the setup of Jackson's experiment does not actually allow for what it stipulates. The result is that Jackson's argument contains a false premise and, as such, cannot establish its conclusion.

2. The Famous Thought Experiment and its Famous Responses

In what he calls the *knowledge argument* against physicalism, Jackson lays out his famous thought experiment and introduces us to Mary.

> Mary is a brilliant scientist who is, for whatever reason, forced to investigate the world from a black-and-white room via a black-and-white

> television monitor. She specialises in the neurophysiology of vision and acquires, let us suppose, all the physical information there is to obtain about what goes on when we see ripe tomatoes, or the sky, and use terms like 'red,' 'blue,' and so on. She discovers, for example, just which wave-length combinations from the sky stimulate the retina, and exactly how this produces via the central nervous system the contraction of the vocal chords and expulsion of air from the lungs that results in the uttering of the sentence 'The sky is blue'. (It can hardly be denied that it is in principle possible to obtain all this physical information from black-and-white television, otherwise the Open University would *of necessity* need to use color television.)
>
> What will happen when Mary is released from her black-and-white room or is given a color television monitor? Will she *learn* anything or not? It seems just obvious that she will learn something about the world and our visual experience of it. But then it is inescapable that her previous knowledge was incomplete. But she had *all* the physical information. *Ergo* there is more to have than that, and physicalism is false. (Jackson 42-3)

Though Jackson avoids elaborate detail, we are to understand the Mary experiment as prohibiting any experience of color whatsoever. Mary has to wear long sleeves and black or white gloves, cannot look at herself in a mirror, cannot press on her eyelids to create phosphene experiences, etc. She can learn anything a (black-and-white) book can tell her, but she may not do anything to gain actual color experience. Anything Mary can learn in her room must be expressible in text, so everything she learns there is factual in nature. In light of this truth, we can interpret Jackson's use of "physical information" to mean "physical facts." For his argument to succeed, then, Mary must come to know every extant physical fact while in her black-and-white room, but, nevertheless, learn a new fact when she leaves it and sees color for the first time. If the above condition is met, since Mary already knew all of the *physical* facts before leaving her room, the new fact cannot be physical, and physicalism must therefore be false. To simplify things, Jackson's argument can be represented in the following manner:

P1 Mary has all of the physical facts regarding color vision (CV) in her black-and-white room.

P2 If physicalism is true, the set of all *physical* facts about CV is identical to the set of all facts about CV.

P3 (a) Mary learns something new when she leaves her room and sees color for the first time.

P3 (b) What Mary learns is a new fact about CV.

C1 Therefore, the set of all *physical* facts about CV is not identical to the set of all facts about CV. [from P1, P3(a), P3(b)]

C2 Therefore, physicalism is false. [from P2, C1]

If we grant both that Mary learns some new fact as a result of her first color experience and that, in her black-and-white room, Mary had all of the physical information, Jackson believes it follows that Mary learns something about qualia, that qualia are nonphysical, and that physicalism is therefore false. Additionally, Jackson takes epiphenomenalism, or the thesis that qualia have no physical effects, to be a consequence of this conclusion (Jackson 45-9). If nonphysical things exist, it is implausible that they could affect physical bodies as this would require the paradox of nonphysical entities having physical effects. Epiphenomenalism solves this problem.

Though brief, Jackson's knowledge argument is powerful and has had considerable influence in contemporary philosophy. In order to combat its conclusion, physicalist philosophers of mind have tried numerous strategies including: 1) Denying that Mary learns anything (rejecting P3 (a)); 2) Admitting that Mary learns something but denying that what she learns is a new *fact* (rejecting P3(b)); and 3) Denying that Jackson's experiment was actually set up to allow Mary to gain all of the physical facts in the first place (rejecting P1). In this section, I will explore traditional uses of these three strategies and evaluate their merits. Finding none of

them fully sufficient, I will draw from their mistakes in order to inform a set of criteria which any successful theory of mind must satisfy, paving the way for my own proposal.

2.1 Denial of New Knowledge

When Mary leaves her black-and-white room, Jackson thinks it "seems just obvious that she will learn something" (Jackson 42). Daniel Dennett disagrees, however, and in "'Epiphenomenal' Qualia?" he argues that Jackson fails to adequately establish this premise, given the critical role it plays in his argument.

Dennett notes that Jackson's appeal to obviousness is actually his only defense of the claim that Mary learns something upon first seeing color. Though Jackson's prediction about Mary is intuitive (and widely shared), Dennett attempts to undermine the claim that it is actually *obvious* by continuing the Mary story with a new twist. In Dennett's continuation, Mary's captors seek to play a prank on her by presenting her with a blue banana for her first color experience (Dennett 60). Presumably, they are attempting to trick her into thinking that the word "yellow" actually refers to the color blue. Surprisingly, however, she is not fooled! Instead, she points out the trickery, proclaiming that the banana she sees is not yellow, like it should be, but blue instead. When her captors inquire as to how she saw through their deceit, Mary replies:

> You have to remember that I know *everything*—absolutely everything—that could ever be known about the physical causes and effects of color vision. So of course before you brought the banana in, I had already written down, in exquisite detail, exactly what physical impression a yellow object or a blue object (or a green object, etc.) would make on my nervous system. So I already knew exactly what *thoughts* I would have (because, after all, the 'mere disposition' to think about this or that is not [a quale], is it?). (Dennett 60).

Dennett's reply on behalf of Mary *is* counterintuitive, but this is not something he denies. Instead, Dennett asks, can it be proven that his response is impossible? Dennett contends that it

cannot, and he explains that the only reason it seems to some like there must be *some* way it can is that human imagination is limited in its capacity (Dennett 61). It is impossible for us in our present state, Dennett argues, to imagine what it would be like to know everything physical, and, as such, we cannot possibly know what it is that a person with unlimited physical knowledge would actually know (Dennett 61). It is, therefore, an open question after Jackson tells Mary's story as to whether Mary learns anything new when she sees color for the first time.[1] This question is begged when Jackson simply asserts that its answer is obvious, and the knowledge argument therefore fails.

Dennett's move, though clever, does not ultimately work. In his continuation of Mary's story, he confuses knowing the thoughts one will have with knowing what one is subjectively experiencing.

Howard Robinson notes the key problem with Dennett's argument: "Knowledge of how someone is disposed to react, verbally or otherwise, does not tell you what it is like to possess a mental state" (Robinson 71). In Dennett's story, Mary claims to know that she has been tricked because she knows things like "seeing a yellow object will cause brain state ABC" and "my brain is in state XYZ," allowing her to infer that she does not see yellow. Additionally, she knows things like "seeing a blue object puts my brain in state XYZ," allowing her to conclude that she is seeing blue.

Knowledge of which brain states accompany sights of each color seems to be what Dennett must mean when he says that Mary will know "the physical causes and effects of color vision" (Dennett 60). Insofar as he believes Mary can predict which thoughts will accompany a

[1] Dennett goes on to argue that epiphenomenalism is flawed and that, since epiphenomenalism is a consequence of Jackson's way of telling Mary's story, Dennett's own way of telling the story is superior (Dennett 67). I will not discuss this line of argument here, however, as this paper's focus is on whether the Mary experiment itself is flawed and not whether epiphenomenalism is coherent.

given color experience, I presume he means by "thoughts" things along the lines of "this rose is red." Insofar as this is all the case, however, Dennett's extension of the Mary story fails to rule out that she learns something new. Knowing only which brain states will obtain when one sees a red rose and which activities (e.g. mental, verbal, etc.) will accompany those brain states does nothing to tell one what having such brain states is actually like from the inside.

Dennett could respond by saying that he meant something more by "thought" than what I am attributing, such that the concept is meant to include thoughts about subjective experience. He might contend that Mary, before leaving her black-and-white room, will know that seeing red for the first time will cause her to have thoughts like "*This* is what seeing red is like!" This response will provide little traction. Dennett specifically denies that Mary's pre-color-experience knowledge includes qualia, and Mary cannot, therefore, know what seeing color is like (Dennett 60). If Dennett meant to include subjective experience in his concept of "thought," then it is clear that he intended Mary to have knowledge only of *which* thoughts would occur upon seeing color and not knowledge of the thoughts' content. At best, Mary can know only that she *will know*, after seeing a color, what seeing that color is like. She can know that she will think "*This* is what seeing red is like," but she cannot actually know what "this" refers to until she leaves her room. Whichever definition of "thought" Dennett employs, Mary stands to learn something new when she sees color for the first time.

2.2 Ability and Acquaintance Hypotheses

Most philosophers do not believe, as Dennett does, that one can deny Mary learns *anything at all* when she leaves her room for the first time. As such, many physicalists have sought ways of admitting that Mary acquires new knowledge when she sees color for the first

time without having to concede that such knowledge is nonphysical. Paul Churchland[2] and David Lewis have, respectively, put forward hypotheses that knowledge of qualia is a kind of knowledge classified as acquaintance knowledge and that knowledge of qualia boils down to a collection of abilities. The goal is to give a physicalist account of qualia which remains consistent with the intuition that Mary learns something upon leaving her black-and-white room.

In "Knowing Qualia," Churchland goes into scientific detail regarding how one becomes acquainted with qualia. Churchland explains that there are several changes which occur in any color-competent organism's nerve pathways and brain upon seeing a color such that the organism becomes able to discriminate for that color in the future (in other words, such that the organism becomes acquainted with the color) (Churchland 166). One might say that once an organism has undergone the requisite neural transformation, that organism "knows" what it is like to see a given color. Of course, this knowledge is not factual; nevertheless, it is the result of purely physical processes and should therefore count as physical itself (Churchland 166).

Churchland does not believe that the term "acquaintance knowledge" picks out a *specific* kind of knowledge, but rather that it describes a more general means for nonfactual understanding (Churchland 165). Lewis supports a similar position in "What Experience Teaches" in that he contends experience knowledge is nonfactual yet physical. Instead of arguing, however, that experience knowledge is an umbrella term for various sorts of nonfactual knowledge, Lewis argues instead that it is a specific set of *abilities*. For Lewis, knowledge of an experience E is identical to a collection of abilities to do things like mentally simulate E, recognize experiences similar to E in the future, etc. (Lewis 97-9). For Lewis, experience knowledge is knowledge of *how* to perform a set of specific tasks, and this is distinct from factual knowledge. A world champion gymnast, for example, knows how to remain on a balance

[2] I here analyze Churchland's take on a view made famous by Earl Conee.

beam while doing flips, but she surely could not provide a list of facts which, once read and understood, would give a non-gymnast her specialized know-how. Lewis holds that knowledge of experience is like the gymnast's knowledge, and, like Churchland, Lewis believes that what Mary learns when she leaves her room is purely nonfactual.

Whether one sides with Churchland that knowledge of qualia is acquaintance knowledge or with Lewis that it is a set of abilities, both replies suggest that Jackson's argument fails to address what Mary actually learns when she first sees color. Because Jackson's premises refer only to facts, it is a problem for him if Mary's new knowledge is purely nonfactual. Jackson's thought experiment was never set up for Mary to have all of the nonfactual physical knowledge, perhaps because Jackson does not believe that there is such a thing. Nevertheless, if Churchland or Lewis is correct, nonfactual knowledge is the only kind Mary learns upon first seeing color, and experiencing qualia teaches the experiencer nothing factual. If either Churchland or Lewis is correct, then, Jackson's experiment is moot.

Knowing the implications that both the acquaintance and ability hypotheses have for Jackson's position, the question now arises whether we should in fact accept either one. I contend that we should not, for both the acquaintance and ability hypotheses are inadequate accounts of what it is to know a quale.

On the ability and acquaintance hypotheses, knowledge of qualia is not "knowledge that" but "knowledge how" or some other kind of knowledge (Churchland 165). The problem, however, is that "knowledge that" seems to better describe what knowledge of experience is. Generally speaking, experience knowledge is understood as knowledge of what some experience is *like*. In order to capture this, however, experience knowledge must be knowledge *that* an experience is like X, where X is a placeholder for the ineffable subjective qualities of experience.

It is true to say that some experience is like X while it is false to say that the same experience is not like X, and, because both of these statements about X have truth values, they must be factual. If the ability or acquaintance hypothesis is true, however, experience knowledge cannot be factual knowledge *that* anything. This is problematic, for to say that experience knowledge does not refer to "what-it-is-like-ness" would be to so dramatically redefine the concept of qualia that it would be unrecognizably different from the concept employed by most philosophical literature. Whether or not the content of experience knowledge is physical, it seems necessarily factual, and neither the ability nor the acquaintance hypothesis can deal with this necessity.

2.3 Equivocation on "Physical"

Given both that Mary learns something when she first sees color and that what she learns is factual, the knowledge argument seems to be gaining ground on the physicalist. Terence Horgan and Daniel Stoljar believe, however, that both of these criteria can be met and the knowledge argument's conclusion refuted if they attack the Mary experiment's structure, denying that it actually allows for what it stipulates.

The Mary case claims that Mary has all of the physical information before she experiences color. Horgan and Stoljar have noted[3], however, that this statement has two possible meanings which will be elaborated upon below. On one interpretation, Jackson is certainly correct that Mary both has all the physical information and lacks other information about the world. On the other, however, Horgan and Stoljar contest that P1 from the beginning of this section is false; Mary learns something when she sees color for the first time only because her permitted methods for gathering physical information (i.e. scientific inquiry) do not actually

[3] See Horgan, Terence. "Jackson on Physical Information and Qualia." and Stoljar, Daniel. "Two Conceptions of the Physical."

allow her to obtain all information that is physical. In discussing Horgan's and Stoljar's very similar positions I will refer to Stoljar and make use of his terminology, but it should be noted that my comments will apply equally to the analogous arguments and terminology put forward by Horgan.[4]

Stoljar distinguishes between two kinds of physicalism: theory based, or t-physicalism, and object-based, or o-physicalism. Some property is t-physical just in case it is either "the sort of property that physical theory tells us about *or* else is a property which metaphysically (or logically) supervenes on the sort of property that physical theory tells us about" (Stoljar 312). In other words, if we can come to know about a property by studying physics and its implications, that property is t-physical.

The set of o-physical properties is more expansive. A property is o-physical just in case it is either "the sort of property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents *or* else is a property which metaphysically (or logically) supervenes on the sort of property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents" (Stoljar 312). The latter disjunct in this definition is just a summary of t-physicalism, but, Stoljar argues, physical science is unable to tell us everything about our physical world, and the first disjunct serves to capture what the second leaves out (Stoljar 313-4).

Stoljar notes that a property can be either dispositional or categorical. If it is dispositional, it depends on other properties for its existence whereas, if it is categorical, it exists independently of anything else (Stoljar 313-4). According to Stoljar, science is capable only of telling us about dispositional properties, and, as evidence, he cites its failure to capture any non-

[4] Where applicable; Stoljar's argument goes into greater depth, so there will be some claims attributed to Stoljar which have no explicit analog in Horgan's argument.

dispositional properties to date (Stoljar 313). Even properties which we might at first think to be categorical, such as extension and mass, in fact depend on other properties themselves; for example, extension is only meaningful if there is a property to define its boundaries, and mass is knowable only by its effects on other things (Stoljar 313). Insofar as we believe it necessary that that each dispositional property be based, somewhere down the line, on a categorical property, categorical properties *do* exist, but science will never tell us about them. T-physicalism holds as physical only those things which science can tell us about, and categorical properties of physical objects are, therefore, non-t-physical. In the definition of o-physical properties, it is these non-t-physical, categorical properties of physical things which the first disjunct intends to include.

Given the distinction above, when we consider what counts as physical in this world we can mean by "physical" either "t-physical" or "o-physical," but both physicalists and epiphenomenalists will have to choose the latter. Physicalists obviously cannot allow that things exist which are nonphysical, but on t-physicalism this is just what must be done with categorical properties. Epiphenomenalists, on the other hand, allow for the existence of nonphysical things but hold that such things must exist without effect on the physical world. On a t-physical view, however, the fact that everything physical must supervene on some categorical property or other (and therefore on something non-t-physical) would contradict the epiphenomenalist's position. Both the physicalist and the epiphenomenalist will find t-physicalism to be unacceptable, and conclusions about the mind which are limited to the t-physical realm should therefore be considered irrelevant to the debate at hand. For both sides of the Mary debate, arguments must be made in the context of o-physicalism.

Turning back to Mary, whose knowledge within the black-and-white room is limited by science, it seems that she can have complete physical knowledge only in the t-physical sense.

This, of course, is by design; knowledge beyond the t-physical would pose a serious threat to the intuition that Mary learns something new. As we have just established above, however, this kind of limited epistemic state is insufficient for Jackson draw any interesting philosophical conclusions. Without complete o-physical knowledge, Jackson's thought experiment has no bite; any facts Mary learns can be dismissed by physicalists as o-physical and Jackson will have shown nothing. Further, as mentioned above, a t-physicalist Jackson will have an irreparably inconsistent position owing to his epiphenomenalism.

Of course, the knowledge argument does not take itself to be an inconsistent way of reaching a philosophically uninteresting conclusion. This, however, means that the knowledge argument must be trying to draw conclusions about o-physicalism instead of t-physicalism, and of this it is incapable. Jackson's setup for the Mary experiment makes clear that Mary can acquire only t-physical knowledge, and it is an invalid move to use "physical" to mean t-physicalism in one's premises but o-physicalism in one's conclusion. For Stoljar, this unacceptable equivocation renders Jackson's thought experiment moot.

Though perhaps compelling initially, Stoljar's argument is inconclusive and at odds with traditional physicalism. In establishing that science cannot tell us about categorical properties—not that it has not done so yet, but that it *can never* do so—Stoljar seems to rely solely on science's past shortcomings (Stoljar 313). Though he calls into question the widely held notion that we *have* discovered categorical properties, he offers no principled reason for thinking that we will *never* do so. There are, of course, principled reasons for thinking science will never discover certain things, particularly those which are logically impossible. There is no justification given, however, for thinking that such logical barriers exist in the case of categorical properties, and without any positive reason for thinking that categorical properties are

undiscoverable in principle, Stoljar's argument entails absurdity. Every discovery in science, after all, has been preceded by a time in which the discovery had not yet been made, and Stoljar's logic does not distinguish between such historical time periods and the current situation regarding categorical properties. If we are to take Stoljar's reasoning seriously, it would have been proper during each historical period of ignorance to have insisted that the discoveries which are familiar today would never come about. While the absurd conclusions entailed by Stoljar's position might be enough to consider rejecting it outright, Stoljar's problems go deeper still.

Stoljar places substantial weight on the t-physical/o-physical distinction, but this distinction cannot do the work Stoljar requires of it. It is insufficient in order to refute Jackson's ultimate conclusion, that qualia are nonphysical, to show simply that Mary was missing physical information while in her black-and-white room. It must also be shown that the missing physical information *was* qualia.

In Stoljar's case, the missing information is information about categorical properties. According to conventional physicalism, however, the things Jackson refers to as qualia are reducible to certain sequences of neuronal firings, and these firings must supervene upon other things (like smaller particles, the laws of nature, etc.). If this conception of physicalism is true, then the traditional physicalist has no reason not to dig in and maintain that qualia are t-physical. The traditional physicalist can say that while Stoljar has made a legitimate distinction between t- and o-physicalism, this distinction does no work in the Mary case because phenomenal information is not o-physical.

If disagreement on the part of fellow physicalists is not bad enough, it is also the case that epiphenomenalists can gain immunity from Stoljar's physicalism by arguing that while Mary lacked o-physical information, this information was not about qualia and therefore has no

bearing on the thought experiment's original efficacy! While the existence of anti-Stoljar strategies for traditional physicalists and epiphenomenalist dualists does not show that they are correct, it does show that Stoljar's inference is invalid; his t-/o-physical distinction can be true while his conclusions about physical reality are false.

Stoljar may have disproven Jackson's claim that Mary acquired all of the physical information, but he does nothing to show that the information Mary lacked is capable of accounting for qualia. In fact, on a traditional conception of physicalism, the information Mary lacks *cannot* account for qualia. Stoljar's strategy, like those before it, is insufficient to establish a flaw with Jackson's central thesis.

2.4 Conclusions so Far

Several unsuccessful positions which attempt to deny that Mary learns a new, non-physical fact when she leaves her black-and-white room have been examined. We have seen Dennett's attempt to deny that Mary learns anything at all upon seeing color for the first time, but this was implausible. The ability and acquaintance hypotheses tried admitting that Mary learns something new but denying that what Mary learns is a *fact* of any sort, and this too was unsatisfactory. Finally, we have witnessed Stoljar's ultimately unfruitful distinction between t-physical and o-physical properties. What we can learn from all of the mistakes thus far is that it looks like a successful physicalist theory of mind needs to a) admit that Mary learns something new, b) admit that what she learns is a fact, c) provide a principled reason for why Mary had no access to this fact in her black-and-white room, d) prove that Mary's new fact accounts for experience knowledge, and e) explain why this fact ought to rightly be described as physical. In the following section, I will go on to present a position which satisfies each of these criteria.

3. A New Approach

Any complete view holding that Mary learns a new fact when she sees color for the first time needs to determine the content of this fact, and my own view is no exception. The fact in question must be sufficient to fully bridge the epistemological gap between pre- and post-color-seeing Mary, and as a result many philosophers have asked what black-and-white Mary is missing that color-seeing Mary is not. For those who admit that the question has an answer (e.g. not those like Dennett, who hold that Mary learns nothing), the near universal response is that Mary is missing experience knowledge.

From Jackson's description of the Mary case and the assumption that Mary learns a new fact, we can infer that the experience knowledge lacked by black-and-white Mary does not consist in anything learnable through the reading of words or the hearing of lectures. After all, Mary is permitted to read anything (in black-and-white) and listen to anything she deems necessary to learning about color, and yet our intuition is nevertheless that she learns something new when she leaves her room. If we are to believe that Mary learns a new fact which cannot be expressed by the English language, of what nature is this fact?

3.1 The Subjective

Thomas Nagel provides us with a plausible candidate for what Mary gains upon leaving her black-and-white room: subjective knowledge. According to Nagel, when science has done all of the explaining it can do we are still left without any understanding of experience's subjective nature (Nagel 14-17). Nagel begins by noting that scientific knowledge, which he calls objective, is a kind of knowledge which can exist independent of points of view (Nagel 14).

Though it is not entirely clear what Nagel has in mind by "points of view," he seems to mean something like "kinds of existence," in the way that I might say I have a typical (i.e. not visually or hearing impaired, etc.) human's kind of existence. Ultimately, Nagel looks to be saying that objective, scientific facts are facts which could be understood by any intelligent forms of life, no matter how unlike human beings they are. There would be no trouble, for example, in explaining to a race of intelligent but deaf Martians the effects that sound waves have upon the inner ear of hearing creatures (provided, of course, that we could work out a mutually understood system of writing or other means of non-oral communication).

Nagel argues that part of what it is to know an *experience*, as opposed to knowing some scientific fact, is inextricably tied to the point of view of the experiencer (Nagel 15-16). It would be impossible, in other words, to explain to the deaf Martian what a trombone sounds like because the deaf Martian cannot adopt the viewpoint of a hearing being. Scientific facts are by nature objective (i.e. viewpoint neutral) and certain facts about experience seem beyond the reach of any viewpoint neutral description. In other words, science cannot fully describe experience because experience knowledge is constituted by, among other things, an aspect which is fundamentally subjective.

Nagel presumably takes the subjectivity of experience knowledge to imply that physicalism is false. He holds that "if we are parts of the world as it is in itself, then we ought to be able to include ourselves—our minds as well as our bodies—in a conception that is not tied exclusively to our own point of view" (Nagel 17). Given Nagel's previous analysis, this is clearly an impossible feat. Our minds cannot be wholly extricated from our viewpoints, and it follows that our minds cannot therefore be "parts of the world as it is in itself" (Nagel 17). If our

minds are not objective parts of reality, however, then what are they? Whatever the answer to this question may be, it looks like our minds can*not* be physical.

3.2 An Attempt at Naturalizing the Subjective

As intuitive as Nagel's premises are and as straightforwardly his conclusion seems to follow from them, Uriah Kriegel offers the physicalist a means of escape in "Naturalizing Subjective Character." Though Kriegel admits that experience has a subjective quality to it (and would presumably admit that Mary lacks knowledge of this quality in her room), Kriegel denies that mental subjectivity need imply that anything exists beyond the realm of the physical. While I will not end up agreeing with Kriegel's position entirely, I think we can learn from it in such a way as will be useful in developing a solid physicalist, naturalist conception of the subjective.

At the outset, Kriegel begins by distinguishing between two components of experience. When a person S sees blue at time *t*, knowledge of S's experience might be described as knowledge of the bluish way it is like for S at *t*, and Kriegel takes the "bluish way" part of this description, which he calls qualitative, to be distinct from the "for S" part, which he calls subjective (Kriegel 23-4). To motivate the important difference present in his distinction, Kriegel notes that "not only *is* the experience bluish, but [the experiencer is] also *aware* of its being bluish" (Kriegel 23). A subject could, in other words, theoretically have a bluish experience without any awareness of the blueness.

In his paper, Kriegel does not discuss the qualitative aspect of experience, but I will address this issue in the following section. Instead, Kriegel focuses on the subjective aspect of experience, hoping to show that it can be naturalized in a manner compatible with physicalism.

Kriegel ultimately finds the subjective component of experience to have two parts—a representation R and a representation of R—neither of which is subjective by itself (Kriegel 46). When a person sees a red tomato, the corresponding experience consists in both a representation of the tomato and a representation of *that* representation. While neither of the constituent representations taken alone are subjective mental states, Kriegel maintains that they are integrated into a single mental state through natural brain processes, and that this resultant state *is* subjective (Kriegel 47). It is not that both R and R's representation exist simultaneously like notes in a musical chord, but rather that R and R's representation have ceased to exist, being replaced by a third mental state which is a complete integration of the first two (Kriegel 48).

In order for a mental state to be subjective, Kriegel contends that it must feature inner awareness, presumably because inner awareness is what we take to constitute subjectivity in the first place (Kriegel 38). In order for a state to fulfill this requirement, it must be capable of self representation, and this is precisely what the integration of R and R's representation accomplishes (Kriegel 38, 46). For Kriegel, when a mental state is both the act and object of awareness, it is properly deemed subjective (Kriegel 38).

While Kriegel's account of what makes a mental state subjective might seem satisfactory, it is unnecessarily complex. A mental state's self-representation is certainly sufficient to establish its subjectivity, but, I will argue, it is not necessary. In the end, Kriegel is wrong that R's representation is non-subjective, and I will show that all that is necessary for subjectivity are higher order representations.

As mentioned before, inner awareness is key to subjectivity, and this is true not only for Kriegel but for Nagel as well—after all, without inner awareness, what is a point of view? Kriegel seems to believe that inner awareness requires self-representational mental states, but is

far from clear why this is so. As long as mental state X is represented to S by mental state Y, it is unclear how it could be the case that S is not subjectively aware of X. Further, S's awareness of X seems in no way affected by whether X and Y are identical or not; an example will demonstrate.

Suppose S has a mental state which is constituted both by a representation of red and a thought that "I, S, am seeing red right now and this is what it is like" (i.e. a representation of that representation). S, in such a case, is surely subjectively aware of redness. It seems, however, that S is equally aware of redness if S's first and second order representations of red take place in two distinct but concurrent mental states. In fact, to take things a step further, it is not even necessary to have the first order representation at all. Let us imagine that S has normal color vision but undergoes some kind of super-neurosurgery. This surgery re-wires S's brain so that, in addition to forming second order representations about red when red is represented to S, S's brain also forms these representations when the sound of the word "red" is represented to S. After hearing the word "red," S believes things like "I, S, am looking at red right now, and *this* is what it is like." Of course, within a few seconds (or maybe even sooner), S likely forms a representation about S's representations of red to the effect of "Wait…these representations are inaccurate!" Nevertheless, there was at least an instant *i* in which S believed that S was seeing red and that S's experience at *i* is what it is like to S for S to see red. At instant *i*, S must be said to have the subjective experience (understood as separate from the qualitative experience) of seeing red. It is, therefore, possible for a person to have a subjective experience by having representation *of* a representation R without actually needing R itself.

At this point, it is natural to ask what the "this" in S's belief that "*this* is what it is like to see red" might refer to if there is no first order representation present. In one sense, "this" does

not refer to anything at all; its referent is a mental state which does not exist. When S possesses second order representations of red, they refer to first order representations of red which S does not realize are absent until the second order representations have already been formed. It is much like when one touches a stove and, believing oneself to be feeling heat and pain, yanks one's hand away only to realize that there was neither heat nor pain during the time one's hand was on the stove.

In the stove example above, the first instant of contact seems to be one in which the person has all of the second order representations—but none of the first order representations—appropriate for having touched a hot stove. There is a feeling that one is experiencing painful hotness, but there is no actual feeling of painful hotness. Experiences like this are, of course, rare in real life and probably never occur naturally in the case of color vision. Nevertheless, there is no logical prohibition against their occurrence, and it is possible—at least in theory—that S should have a subjective (again, understood as distinct from the qualitative) experience of red without actually having red represented to S. Insofar as the analysis to this point has been correct, subjectivity requires only higher second order representations.

3.3 Why the Qualitative is Irrelevant to Jackson's Thought Experiment

Having naturalized the subjective, one may still feel that a problem is posed for the physicalist by the qualitative aspect of experience, and before we can delve into how our findings thus far affect the Mary case, this worry must be addressed. I will argue that factual knowledge about any given representation is necessarily a higher order representation itself. Jackson's experiment allows Mary to possess only the physical *facts*, and, if first order representations

cannot be factual, Jackson's experiment says nothing about physicalism if Mary lacks only first order representations in her black-and-white room.

To start, I believe that the prevailing belief that first and second order representations are essentially the same things as first and second order thoughts is mistaken. The most basic way in which one can *think* about something is not the most basic way in which something can be *represented* to an individual. First order representations, as I understand them, are difficult to describe, for they are not rightly expressible through language. When I look at a red circle, my first order representation of what I am seeing is not "I am seeing a red circle." This is, perhaps, the most basic linguistic representation of my experience, and it is a first order thought, but it is not, ultimately, the most basic way in which my experience is represented to me. The best way I can describe the simplest, and therefore the first order, representation of a red circle to its observer is through analogy to a photograph. When I see a red circle, it is as if my eyes are taking a picture of the red circle and presenting it to my mind. The presentation of said picture, being the most primitive means by which a visual image can be represented to my mind, is a first order representation.

Before any thoughts are formed about it, an image before my mind is only that, a raw image. The image can be described with words, but these words only describe the image and do not constitute it. In fact, to verbally describe a photograph is to represent it, and a verbal description of a first order representation is a second order representation. If this is not intuitive, an example more complex than a red circle will illustrate.

Suppose I was mugged yesterday, and I want to file a police report. Presumably, I have a mental image of my assailant that I will describe to the police; this image is my first order representation of the mugger's appearance. I will tell the police things like the attacker's sex,

race, height, weight, hair color, etc. Each of these bits of information represents a little bit of my mental image and, assuming I could give a perfect description, they would represent the entire mental image when taken together. This representation of my mental image, however, is just that: a representation. It would be incorrect to say that the lengthy description just *was* my first order representation in a different form. A compilation of sentences is not at all like a picture. The compilation can *represent* a picture, but it cannot be identical to one.

Let us now extrapolate from the mugging case. In no instance of vision is one's original representation of the external world in language—even if one looks at a sentence, the relevant first order representation is a picture of the sentence, not the sentence itself. When I feel something rough, I have a tactile snapshot of roughness; this is my first order representation in such a case. When I hear a sound, I have an auditory snapshot; the same goes for the rest of the senses. It quickly becomes clear that one cannot have a linguistic first order representation of the external world.[5] Once raw sense data is described by language, it is represented, not duplicated, and one can no longer be in the first order of representation.

Linguistic content is not all that first order representations lack. It seems clear that representations like my image of a red circle, before any thoughts respecting them are formed, are intentional only with respect to the things in the real world which they represent. Concepts, of course, are not things in the real world and cannot be called before one's mind *qua* concepts without bringing new thoughts along with them. If I think of the concept of a circle, for instance, I think of the defining feature of a circle: roundness. Once I see a circle and think "circle," I have implicitly thought "this object has conformed to my concept of circle because it is round" and have, in the second order, represented my mental image of a circle as something which

[5] Things get more complicated if one moves beyond the senses into the realm of introspection, but I will not deal with such complications here.

conforms to my concept of circle. First order representations can therefore make no reference to mental concepts, for to do so would be to introduce extraneous thoughts and leave the first order of representation. Representations of the first order must be both nonlinguistic and nonconceptual.

Factual knowledge, by contrast *is* conceptual. The factual knowledge that *P*, for instance, always involves the concept of *P*'s object *O*. As a result, representations containing *P* cannot be first order; they will always be representations *of* a representation of *O*. The kind of knowledge expressed through factual statements is "knowledge that…" Once the ellipsis is replaced with a proposition, conceptual content is inevitable. This content rules out the possibility of any "knowledge that…" (i.e. factual knowledge) being a first order representation, and facts, therefore, must always be representations of a higher order.

Having established the impossibility of a factual first order representation, the implication for Jackson's thought experiment is clear. In section 2, P3 (b) states that Mary learns a new fact about color vision. If all that Mary gains upon seeing color for the first time, however, is a first order representation, then nothing she learns can be factual. If Jackson cannot claim that Mary learns about more than qualitative aspect of experience, then he is stuck with a false premise and can say nothing about physicalism.

3.4 Implications for the Mary Case

Having answered the worry about phenomenal aspects of experience, we can now apply our insights thus far to Jackson's thought experiment. At the beginning of section 3, I began by asking what it is that Mary lacks in her room and gains upon seeing color for the first time; we are now in a position to answer that question. We have already established that black-and-white

Mary lacks subjective knowledge, but that only prompted us to ask what the subjective consists in. From there, we arrived in our present position, having noted that higher order representations constitute subjectivity. Subjective knowledge, then, is just knowledge of these higher order representations. Now, however, we have one final question to answer: What does all of this mean for physicalism?

We know that what Mary lacked in her room was a collection of higher order representations of color. What we still need to determine, however, is whether these representations can be said to constitute i) facts and ii) physical facts. If the answer to i) is "yes" while the answer to ii) is "no," then the Mary experiment succeeds, having shown that, upon leaving her black-and-white room: Mary learns something new, what she learns is a fact, and the fact she learns is not physical. If, however, the answer to both i) and ii) is "yes," then Jackson's experiment fails, and we will have proven that P1 from section 2 is false—the case did not actually allow Mary access to all of the physical facts before seeing color.

The answer to i) is "yes." The higher order representations relevant to Mary's case are thoughts like "I am seeing red right now" and "*This* is what it is like to see red!", both of which are clearly factual statements—they can either be true or false. The answer to ii) is also "yes." To see why this is the case, we need only note that thoughts which constitute second order representations are, in fact, *thoughts*. As mentioned earlier, the proponent of Jackson's thought experiment is an epiphenomenalist, and epiphenomenalists admit that thoughts are physical. Higher order representations, and therefore the subjective information which Mary lacks, is factual and physical.

Surely, the proponent of Jackson's experiment might argue, there is something wrong with the analysis above. It is, after all, stipulated that Mary *does* have all of the physical facts in

P1 from section 2. My response to this is simply that P1 is false under the parameters of the Mary experiment. Mary is prohibited from experiencing color, and this means that she cannot have the relevant higher order representations. Surely, she *can* think the sentence "This is what it is like to see red!", but this is not the same thing because, when she thinks such things in her black-and-white room, the thoughts are not genuine. A higher order *representation* of color is not simply a higher order *thought* about color; for a thought to actually be a representation of anything, it must also be believed. A representation of something to one's mind is how one experiences that thing, and if Mary thinks up a sentence about red but does not believe it, then she does not actually experience red in a manner consistent with her thought. In such a case, Mary's thought is not a representation of anything, for it does not express how anything is actually experienced by her. For Mary to have an actual higher order representation of red, then, she must have a thought about red which is consistent with the way she experiences red.

Mary can never gain what she needs in order to form a genuine higher order representation about color (namely, super-neurosurgery or a color experience) while in her black-and-white room. The kind of belief she needs cannot be acquired from black-and-white books or black-and-white screens, and P1 is false. It was never true that Mary had all of the physical information, and Jackson's thought experiment therefore rests upon an unsound argument by reason of a false premise.

4. Objections

Having argued that the Mary experiment fails, I will now examine and reply to a few potential objections. First, I will evaluate the claim that higher order representations do not

provide an exhaustive account of the subjective, and I will proceed by answering worries that higher order views lead to infinite regresses.

4.1 Higher Order Accounts Cannot Account for the Subjective

Subjective experience is, in a way, ineffable. If one were to barge into Mary's room in an attempt to explain what seeing color was like, there would be no words adequate for the task. How is it then possible, a proponent of the Mary case might ask, for representations which have been determined to be propositional in nature (i.e. expressible in language) to fully account for the richness of subjective experience?

While the motivation behind the question above is understandable, it is ill-informed. The objector assumes I am committed to higher order representations being wholly expressible through language, and this is not the case. It is easy to see how a superficial reading of my arguments would lead to this conclusion, but let us delve deeper.

I have said that the experience of seeing red is just certain higher order representations like "I am seeing red right now" and "*This* is what it is like to see red." The former of these representations has only linguistic components, and it is obvious that it alone cannot account for subjectivity. The latter representation, however, while expressed in words on this page, seems to have a clear non-linguistic referent. Specifically, "this" is a placeholder word for the very ineffable qualities of experience mentioned at the beginning of this subsection.

For proponents of the Mary case, my use of a placeholder may look like cheating. I have, they may protest, hidden supernatural, epiphenomenal content in the guise of a physical thought. On such a view, I have merely taken qualia and called them thoughts, but this is not a permissible move. If true, this is indeed quite damning to my position, but it fundamentally rests

upon a false assumption, namely, that a physical thought must be fully expressible in some extant language.

Thought does not fundamentally depend on language. For many this is intuitive, but those for whom it is not should consider the following. There are three possibilities regarding the development of human language and thought: either 1) they arose exactly simultaneously; 2) language came first; or 3) thought came first. The first of these options is highly improbable, as it would require all human beings in the first generation of thinkers to have somehow spontaneously begun thinking in a language which, before the moment of first thought, did not exist. The second option is even less probable. Option 2) stipulates that language developed without anyone thinking about it at all, but, given the complexity of language, this seems far from likely. A coherent language requires planning, and planning requires thought. This leaves us with option 3), that thought developed first. This third option seems to make the most sense, as it implies that early humans had thoughts about things which they wanted to express, and then developed language in order to do so. For example, prehistoric people presumably had certain thoughts which corresponded with a concept of the animal now referred to as woolly mammoth; there were just no words for the concept yet. Thoughts must be conceptual, but they need not be linguistic.

If thought once existed without any language at all, then language, surely, is not required to render thoughts physical. Are we to assume that human thoughts before a certain date—that is, the date on which language was first used—were nonphysical but *became* physical immediately upon their expression in the world's first language? Surely not, but Mary sympathizers can here revise their positions such that it is only necessary for physical thoughts to be expressible through language *in principle*. In other words, there need only be a theoretically

possible language in which physical thoughts are expressible, and any *actual* such language is unnecessary.

This move certainly improves the position of those who support the Mary case, but it will not ultimately be their salvation. Pre-linguistic humans had thoughts which were not in any language's words. While there are probably possible sentences which can fully capture some, or even most, of these thoughts, it is unrealistic to expect that there should be a one-to-one mapping for all physical thoughts. Those who do not share this intuition should take into account that there are words and phrases in certain languages which have no precise translation in others. Given this, it seems highly unlikely that every thought in a mind completely uninhibited by linguistic constraints should be mappable one-to-one upon any hypothetical language. There is no reason to suspect, therefore, that some thoughts which are rightly considered thoughts as opposed to qualia are indescribable. Further, there is nothing which would prevent many of these indescribable thoughts from being the kind which can be true or false, and there is therefore no reason to suppose that ineffability necessarily implies either that a thought is nonfactual or that it is nonphysical. Given all of this, it is perfectly coherent to posit the existence of physical, ineffable facts, and the charge of my masking the nonphysical with physical language fails. Because it is permissible to hold both that the subjective consists in factual thought and that these factual thoughts are ineffable, the original criticism, that higher order thoughts cannot adequately account for the subjective, fails as well.

4.2 Infinite Regress

Another potential objection is that my position leads to an infinite regress. If subjective awareness is just certain higher order representations then, in order to be subjectively aware of

these higher order representations, representations of still higher orders are needed. There is no reason, the objector says, that these representations ever need to run out. Surely however, this is absurd and we do not actually have an infinity of higher order representations.

In response to this kind of objection, I agree that an infinite regress is absurd, but I disagree that my arguments entail one. To be subjectively aware of one mental state is to have a higher order representation of it, and to be aware of that higher order representation is to have another representation of a still higher order. There is no reason to think, however, that these progressively higher order mental states continue, or even could continue, to infinity.

The best way to get confused into believing that my position requires an infinity of higher order representations is to think that subjective awareness of a representation R requires awareness of each representation *of* R. For example, one might think that for Mary to be aware of seeing red, she must also be aware *of* being aware of seeing red. And then to infinity. Fortunately, this just does not seem necessary.

When I hold a pen, I am aware that I am holding that pen. I am usually also aware (either through an occurrent or standing belief) that I am aware of my holding of the pen. But this is where it usually stops; if you asked me whether I was further aware of this awareness, I would have to stop for a minute and figure out what it even meant to have a fourth order representation. I might eventually answer with "yes," but at this point it seems more appropriate to describe my behavior as forming a new belief than it does to say that I merely accessed a standing one. (For example, after figuring out a complicated math question, knowledge of the answer is a new belief and not merely a standing one which has been accessed.)

It seems that I can be subjectively aware of something on a high enough level such that I am not actually aware of that awareness. Further, lacking awareness at one order does not seem

to undermine my awareness at another; in the pen example one would surely not want to say that it turns out I was actually unaware of there being a pen in my hands. Stipulating that subjective awareness requires awareness all the way up, then, does not seem to be a winning strategy.

5. Conclusions

Much has now been considered. Three prominent physicalist theories have each been proven unable to satisfactorily survive the knowledge argument. From their failures, however, we have been able to extract a set of criteria for success and use it to construct a new defense against the Mary case. This defense, at its core, amounts to the theory that subjective consciousness is higher order representation and the charge that, while Jackson stipulates Mary has all of the physical information, he sets the rules of his thought experiment so as to prevent her from actually acquiring some key information that is rightly considered physical. This information, constituted by higher order representations, is able to account for the epiphany Mary has when she leaves her room. In the end, there is no need to posit anything nonphysical, and the knowledge argument does not threaten physicalism.

Works Cited

Churchland, Paul M. “Knowing Qualia: A reply to Jackson (with Postscript: 1997).” *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Ed. Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar. Cambridge, MA: MIT, 2004. Print.

Dennett, Daniel C. ““Epiphenomenal” Qualia?” *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Ed. Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar. Cambridge, MA: MIT, 2004. Print.

Horgan, Terence. “Jackson on Physical Information and Qualia.” *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Ed. Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar. Cambridge, MA: MIT, 2004. Print.

Jackson, Frank. “Epiphenomenal Qualia.” *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Ed. Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar. Cambridge, MA: MIT, 2004. Print.

Kriegel, Uriah. "Naturalizing Subjective Character." *Philosophy and Phenomenological Research* 71.1 (2005): 23-57. *JSTOR*. Web. 15 Oct. 2012.

Lewis, David. “What Experience Teaches.” *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Ed. Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar. Cambridge, MA: MIT, 2004. Print.

Robinson, Howard. “Dennett on the Knowledge Argument.” *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Ed. Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar. Cambridge, MA: MIT, 2004. Print.

Stoljar, Daniel. “Two Conceptions of the Physical.” *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Ed. Peter Ludlow, Yujin Nagasawa, and Daniel Stoljar. Cambridge, MA: MIT, 2004. Print.

Nagel, Thomas. “Mind.” *The View From Nowhere*. Oxford: Oxford UP, 1986. Print.