# Dualistic Physicalism:
# From Phenomenon Dualism to Substance Dualism

Joseph Polanik, JD

# Table of Contents

## *Preface*

In his video presentation to the 5th Online Conference on Consciousness, Daniel C. Dennett acknowledged that nothing in the brain is identical to an afterimage; that dualism follows from affirming the existence of the afterimage; and, that he would defend materialism by denying the existence of the afterimage. After examining the Dennett defense, I will state an argument for dualism from experience which affirms what Dennett denies; but which, without further assumptions no one is required to make, constitutes a form of dualism milder than either property dualism or substance dualism: *phenomenon dualism*.

The essential claim of phenomenon dualism is that there are two distinct kinds of phenomena, physical and experiential.

Recognizing phenomenon dualism as a form of dualism distinct from both property dualism and substance dualism provides a new perspective on the dispute between John R. Searle and his critics as to whether he is a property dualist. I will argue that, with respect to the brain/experience relation, Searle makes no claims that would justify classifying him as either a property dualist or a substance dualist; so, he belongs in a camp other that that of either Chalmers or Descartes.

However, I will also argue that his critics are right to think of Searle as a dualist of some sort; and, further, that Searle is making a stronger claim than either predicate or conceptual dualists; namely, that there are two distinct kinds of phenomena to talk about. In my view that makes him a phenomenon dualist.

Besides supporting a more satisfying classification of Searlean philosophy, the concept of phenomenon dualism developed here has another, more significant consequence: one may reject Dennett's eliminativism concerning first-person phenomenology without embracing a form of dualism that defeats the possibility that a scientific explanation for the occurrence of first-person phenomena will be found … some sweet day in the future.

Next I defend phenomenon dualism against physical/phenomenal identity theory, the second way (after eliminativism) of holding that there is only one kind of phenomena. I do this in the context of discussing Jackson's Knowledge Argument which is easily understood as highlighting the distinction between the physical phenomena relating to color vision that Mary knew about before her release and the experiential phenomena constituting color experience that Mary only becomes acquainted with after her release. I conclude that we have good reasons to affirm the non-identity of physical and experiential phenomena.

After defending phenomenon dualism as to the brain/experience relation, I turn my attention to a more contentious topic, the brain/subject relation.

In common with many philosophers, I hold that the occurrence of experience presupposes the existence of the experiencing subject. Consequently, I reject eliminativism as to the experiencing subject. However, I also argue for brain/subject non-identity, a less common position.

Such a position is already dualistic; but, a philosopher who wants to have a

complete theory of the brain/subject relation must take a position on the question of subject causation: *Am I causally effective at least some of the time?*

Denying that the experiencing I is causally effective at any time constitutes epiphenomenal dualism, which position I reject. I'm not claiming to have proven from unassailable first principles that epiphenomenal dualism is false; only that I reject it for what I believe are good reasons. If I am wrong, if it turns out that I am not casually effective at any time, I am an epiphenomenon who denies being an epiphenomenon.

Affirming that I am causally effective at least some of the time is the defining characteristic of the position I will call *subject causationism*; although, assuming traditional philosophical terminology, it would be a form of interactive substance dualism.

## *§1 The Central Question*

> In the history of philosophy, this has been the center of the traditional mind–body problem: How exactly does consciousness relate to the brain and to the rest of the physical world? (Searle, 2007, 169)

These are the key, rarely disputed facts with which scientists and philosophers of consciousness must concern themselves:

- There is experience.
- There is brain activity.

We've known that these facts are correlated for about 2300 years – ever since that sweet day in antiquity, the Day of First Correlation, when the Greek physician, Herophilus, rejected Aristotle's claim that the brain just sat there cooling the blood and proposed that the brain is the seat of consciousness.

Ever since that day, we've made tremendous progress in understanding the brain and in discovering which brain activity is correlated with which conscious experience. However, despite the passing of centuries, the rise and fall of civilizations, the relentless advance of technology, we've not made much progress in discovering why and how these psychophysical correlations occur in the first place; and, there is no guarantee that we will ever succeed in doing so.

Thus, the unanswered question remains: *How are we to describe and explain the brain/consciousness relation?*

Following Searle, I will leave the task of explanation to scientists[1]. I will focus on describing the problem to be solved; and, I will begin by clarifying the two relevant meanings of “consciousness” as used in the phrase “brain/consciousness relation”.

In one sense, “consciousness” just means “experience”. My stream of consciousness is my stream of experience, the ongoing flow of experiential phenomena that I experience. Consequently, one aspect of the brain/consciousness relation is the brain/experience relation.

In another sense, “consciousness” means subject of experience. I am conscious of experiencing. I am *this* that is consciously experiencing *that* afterimage. It seems as natural to conclude that I am *this* consciousness as it is to conclude that I am *this* subject of experience (or *this* experiencer or *this* experiencing I or *this* whatever). Consequently, another aspect of the brain/consciousness relation is the brain/subject relation.

In my view, one may reasonably expect a theory of the brain/consciousness relation to describe and explain both the brain/experience relation and the brain/subject relation; otherwise, the job's not done. Uriah Kriegel similarly

---

1 “… a large part of the philosophical task is to clarify the problem conceptually … once the problem is cleaned up, the philosophical job is over and the factual empirical issues should be solved by lab scientists”. (Searle, 2007a, 169)

claims that a full theory of consciousness must include two component theories.

> When I have a conscious experience of the blue sky, there is something it is like for me to have the experience. In particular there is a bluish way it is like for me to have it. ... The bluish way it is like for me has two distinguishable components: (i) the *bluish* component and (ii) the *for-me* component. I call the former *qualitative character* and the latter *subjective character*. (Kriegel, 2009, 1)

Kriegel then assumes that a full theory of consciousness "would include accounts of both qualitative character and subjective character" and goes on to develop his component theories separately, a practice that I will follow.

## *§2 The Brain/Experience Relation*

There are any number of ways to characterize the brain/experience relation and innumerable theories that attempt to explain it or argue that it is inexplicable. Decades or even centuries of effort by scientists and philosophers may be required before the mystery is finally solved or found to be unsolvable.

To make any progress at all, one must state the problem to be solved; and, to do that, one must adopt some perspective from which to view the problem. Galen Strawson states the problem in terms of kinds of phenomena, physical phenomena and experiential phenomena, a practice I will follow.

> My problem is the old problem ... what is the relation between the phenomena of conscious experience and physical phenomena? In other words, what is the relation between *experience* and *matter*? (Strawson, 2010, 44 (emphasis in the original)[2])

However, since it is commonly held that a relation presupposes the existence of its relata, a philosopher might conduct a pre-emptive strike against any and every possible theory of the brain/experience relation simply by denying the existence of experience.

Daniel C. Dennett is one such a philosopher. In his video presentation to the 5th Online Conference on Consciousness, Dennett guides the viewers through the process of inducing an afterimage of an American flag; focuses the viewer's attention on one particular red stripe in that afterimage; and, draws an important conclusion based on Leibniz's Law of Identity (the Indiscernibility of Identicals).

> If A is a red stripe and nothing in the brain is a red stripe, then nothing in the brain is identical to A which has to be somewhere else! Dualism follows. In my opinion this is the shortest, sweetest and actually in the end the most convincing argument for dualism I know; and, as a good materialist, I have to resist this. (Dennett, 2013, @12:10)

Dennett argues that the materialist may resist the conclusion of dualism by denying the existence of the red stripe one experiences in a flag afterimage.

> We can talk about "that red stripe". ... We have no trouble referring to that red stripe. ... It's a thing for us, as good as any other thing in our experienced world. It's a part of our experience in that sense. And yet I'm saying it doesn't exist. It doesn't exist. It only seems to us that it exists. (Dennett, 2013, @20:45)

## §2.1 The Elements of Dualism

In essence, Dennett holds that dualism follows from the conjunction of two claims:

1. (*The Existence Claim*) – Experience exists; and,
2. (*The Non-Identity Claim*) – The brain/experience relation is not identity.

This statement of the elements of dualism has considerable support among

---

2 In all quotes with emphasis, the emphasis is in the original unless otherwise specified.

philosophers. For example, in considering whether epiphenomenalism is the price of dualism, Stephen Yablo suggests that a certain minimalistic dualism is widely held.

> Why should epiphenomenalism concern anyone today? Part of the answer is that dualism is not dead, only evolved. Immaterial minds are gone, it is true, but mental phenomena (facts, properties, events) remain. And although the latter are admitted to be physically realized, and physically necessitated, their literal numerical identity with their physical bases is roundly denied. (Yablo, 1992, 246)

In a footnote to this passage, Yablo continued

> In case it seems odd to describe the picture just outlined as dualist, bear in mind that all I mean by the term is that mental and physical phenomena are, contrary to the identity theory, distinct, and contrary to eliminativism, existents. (Yablo, 1992, 246, fn 2)

Jaegwon Kim (1989) is even more succinct, “the choices we face concerning the mind-body problem are rather stark: there are three – dualism, reductionism, and eliminativism”.

Sometimes the question of what constitutes to dualism arises in connection with theories concerning a causal relation between brain and experience. In his critique of McGinn's mysterianism, Uriah Kriegel assumes that a causal relation between brain states and conscious states would be a non-identity relation.

> McGinn appears to assume that conscious states are caused by brain states. His argument does not go through if conscious states are simply identical to brain states. In other words, the argument does not go through unless any identity of conscious states with brain states is rejected. But such rejection amounts to dualism. (Kriegel, 2007a, 38)

Searle appears to concur as well.

> … if brain processes cause consciousness, then it seems to many people that there must be two different things, brain processes as causes and conscious states as effects, and this seems to imply dualism. (Searle, 1997, 7)

Searle doesn't contest the claim that a true causal (cause and effect) relation would be a non-identity relation. Instead he argues for an alternate concept of causation, non-event causation, that would not be dualistic.

While I agree with this view of the elements of dualism, it seems reasonable to acknowledge an intermediate step implicit in any argument that dualism has been achieved: the conjunction of the existence and the non-identity claims is sufficient for ontological irreducibility which is sufficient for dualism.

In any case, nothing in the elements of dualism specifies the *type* of dualism that it detects; and, it is not clear from the rest of Dennett's presentation whether he holds that some particular form of dualism – property dualism, substance dualism or whatever – follows from affirming the existence of experience while denying brain/experience identity.

Until Dennett clarifies his position on that point, I will interpret his claim that “dualism follows” as if it were written “dualism of some sort follows”. If it turns out that when Dennett says “dualism follows” he means that interactive substance dualism or some other specific form of dualism follows from these two

premises, my interpretation is still valid. A claim that dualism of the interactive substance sort follows from the conjunction of the existence claim and the non-identity claim would imply that dualism of some sort follows from those premises; and, similarly for claims that property dualism or some other specific form of dualism followed from those two premises.

## §2.2 Proceeding from Common Ground

Dennett and I agree as to the non-identity claim. We also agree as to the consequence of affirming the existence of experience while denying brain/experience identity – dualism follows. But, here the road forks: we disagree as to the existence claim.

To defend materialism, Dennett opts to deny the existence of the red stripe; so, his defense of materialism consists of adopting eliminative materialism and defending that.

To avoid eliminative materialism, I affirm that the red stripe in my experience while I am experiencing a flag afterimage exists while I am experiencing it. Consequently, I have affirmed a state of affairs from which follows dualism ... *of some sort*.

In what follows, I will compare these choices and their consequences with the aim of showing that dualism is preferable to eliminative materialism.

In §2.2.1 I will examine the consequences of the Dennett defense of materialism. In §2.2.2 I will present my argument for dualism from experience to show *how* dualism follows from the conjunction of the existence claim and the non-identity claim. In §2.2.3 I will answer the question, *what sort of dualism follows from this conjunction?*

I aim to show that, while Dennett is correct to say that dualism follows from affirming the existence of experiential phenomena not identical to physical phenomena, the form of dualism that follows is very mild. In particular, I will argue that it is not obviously incompatible with the possibility of scientific explanation.

### *§2.2.1 Evaluating Dennett's Defense of Materialism*

Dennett's statement that the red stripe doesn't exist is ambiguous. Does it mean that the red stripe in my experience while I am experiencing a flag afterimage does not exist in any sense whatsoever; or, merely that it does not exist in an experiencer independent way?

If Dennett is merely making the latter claim, it is not clear who he thinks he is arguing against. I know of no one who argues that the red stripe in one's experience while one is experiencing a flag afterimage is not identical to anything in the brain *and* exists in an experiencer independent way.

Consequently, I take Dennett to be asserting the former claim. Hence, the premise of the Dennett defense of materialism is rather extreme; but, what is the alternative? If Dennett admits that the red stripe in his experience while he is

experiencing a flag afterimage exists *in any sense* despite not existing as a material object; then, he is admitting to recognizing two modes of existence; precisely the claim that makes Searlean philosophy dualistic.

On the other hand, if Dennett is in fact claiming that the red stripe does not exist in any sense, he has a Defense Gap. His opponents will only grant that the red stripe does not exist as a material object; so, Dennett would still have to show that it does not exist in any sense whatsoever.

### §2.2.1.1 The Contradiction in the Dennett Defense

How is the defender of materialism to proceed? He will deny the existence[3] of the red stripe because "The red stripe is an intentional object which doesn't have to exist" (Dennett, 2013. @17:10).

As it happens, I am willing to stipulate that an intentional object does not necessarily exist *as a material object*; and, further, that the red stripe is an intentional object of just that sort, one that does not exist as a material object.[4]

However, if the red stripe in experience does not exist in any sense whatsoever, how can it *be* an intentional object … or anything at all? Logically, if the red stripe *is* an intentional object, it must *exist as* an intentional object.

Dennett admits that I can refer to the red stripe in my experience while experiencing a flag afterimage; so, I'll refer to it as 'a'.

> **a = the red stripe in my experience while I am experiencing a flag afterimage**

Dennett admits that a is an intentional object; so, letting 'F' represent the predicate 'intentional object', we get

> **[E-1] Fa**

By the Law of Identity, a is itself and not something else.

> **[E-2] a = a**

Naturally, we can combine [E-1] and [E-2] by &I (And Introduction) without additional assumptions.

> **[E-3] a = a & Fa**

Now, the rule of Existential Generalization (aka Existential Introduction,

---

3 Dennett says he will quine the red stripe. "To quine is to deny the existence of something real and important" [@ 15:40]. Calling his method 'quining' seems to imply that Dennett acknowledges that the red stripe in his experience after he induces a flag afterimage is real in some sense … but that he's going to deny its existence anyway. If so, there may be a further objection to the Dennett Defense; namely, that it makes no sense to deny the existence of something one concedes to be both real and important.

4 The wording of Dennett's premise, that the red stripe is an intentional object which doesn't have to exist, seems to suggest that he is making a modal claim; perhaps, something like "An intentional object doesn't necessarily exist". If that is his claim, he doesn't explain how to make the leap from "doesn't necessarily exist" to "actually doesn't exist" which is what he would need to show to make the Dennett Defense of materialism succeed.

http://en.wikipedia.org/wiki/Existential_generalization) allows us to transform [E-3] into [E-4].

**[E-4] (∃x)(x=a & Fx)**

The sequence [E-1] through [E-4] is Theorem 137(a) from Lemmon (1965). [E-4] could be translated into English as “There is something, a, that is an intentional object” or “There exists something which is an intentional object and it is a”.

By definition, existentially quantified propositions make existence claims; so, [E-4] asserts the existence of a on the basis of [E-1]. Consequently, Dennett contradicts himself by asserting [E-1] and denying [E-4] which is derivable from [E-1].

*Does this conclusion seem strange?*

It shouldn't.

Attributing a predicate presupposes the existence of the subject of predication. It can not be the case that 'S is P' is true while simultaneously 'S is' is false. Since, for purposes of logical thought, 'is' means 'exists', saying “The red stripe in my experience while I am experiencing a flag afterimage is an intentional object that does not exist as an intentional object” is simply a senseless self-contradiction.

I am not willing to suspend standard logic to make eliminative materialism seem plausible. That is too high a price to pay in collateral damage.

### §2.2.1.2 Other Problems

In principle, Dennett could simply reverse himself and deny that the red stripe is an intentional object. To eliminate the risk of analogous contradictions, he would also have to refrain from attributing any other predicates to the red stripe. He would have to deny my claim that the red stripe in my experience while I am experiencing a flag afterimage is an experiential phenomenon; and, he would have to deny every other claim of the same form.

Eliminating the source of these contradictions would weaken Dennett's attempt to make his defense of materialism seem plausible; and, he would still face other problems.

#### §2.2.1.2.1 Referring to Non-Existents

Refraining from making claims that attribute predicates to the red stripe as the subject of predication may eliminate the incentive to claim that one may refer to something that does not exist in any sense whatsoever. If not, then Dennett should offer a theory of referencing that permits referring to that which does not exist in any sense whatsoever.

He would also need to explain how we can distinguish one instance of non-existence from another. I can induce a greenish looking afterimage as easily as I can induce a reddish looking afterimage. If neither one of them exists in any sense whatsoever, how do I tell them apart?

### §2.2.1.2.2 Violation of Common Sense

Consider the following statements about a well documented weather phenomenon, the tornado.

> [T1] Tornadoes occur.
>
> [T2] A tornado is occurring right now.

Assuming that Earth is our universe of discourse, [T1] is true at all times, even when there are no tornadoes occurring anywhere on Earth at the time the statement is made. In contrast, [T2] is false when there are no tornadoes anywhere on Earth; but, it is true if at least one tornado exists on Earth at the time the statement is made. Thus, we may conclude that a tornado exists while it is occurring. Generalizing to all phenomena, we may say that a phenomenon is instantiated by its occurrence. A phenomenon exists while it is occurring

An experiential phenomenon such as an afterimage occurs when it occurs to its experiencer; so, it is *instantiated* by its *occurrence* to an *experiencer*.

Now, having guided his audience in inducing an afterimage, Dennett can hardly deny that afterimages occur or that, while the experience in ongoing, an experiencer of an afterimage may truthfully say “An afterimage is occurring to me right now”. So, if Dennett wants to argue that an afterimage doesn't exist while it is occurring, he would have to justify the wildly implausible claim that experiential phenomena are unique in that, unlike physical phenomena, they do not exist in any sense whatsoever even while they are occurring to the subject of the experience.

*Is there a coherent way to do that?* I don't think so.

### §2.2.1.2.3 Denial of Experience

Adopting eliminative materialism to avoid dualism in any form requires the complete denial of experience.

Just as nothing in my brain is identical to the red stripe in my experience when I am experiencing a flag afterimage, nothing in my brain is identical to the red stripe in my experience when I am looking at a physically instantiated American flag. Consequently, Dennett's method of quining a particular aspect of a particular experience would be equally effective against any aspect of any experience. Hence, by denying the existence of the red stripe in his experience while experiencing a flag afterimage, Dennett has effectively denied the existence of experience – *any and all experience*.

Assuming that Dennett would affirm the existence of experience, this would be an obvious self-contradiction. However, it is unlikely that Dennett would affirm the existence of experience. Indeed, all signs are against it. In Consciousness Explained, Dennett denies the existence of first-person phenomenology.

> There seems to be phenomenology. That's a fact that the heterophenomenologist enthusiastically concedes. But it does not follow from this undeniable, universally attested fact that there really is phenomenology. This is the crux. (Dennett, 1991, 366)

Actually, some philosophers argue that, from the fact that there seems to be

phenomenology, it *does* follow that there *is* phenomenology. Strawson makes this point in Mental Reality: "… for there to seem to be rich phenomenology or experience *just is* for there to be such phenomenology or experience" (Strawson, 2010, 52).

Dennett's rationale for thinking that we are mistaken to believe that we are experiencing, say, color qualia is that

> These additions are perfectly real, but they are ... not made of figment, but made of judgment. There is nothing more to phenomenology than that. [Dennett, 1991, 366]

While elaborating his theory of heterophenomenology, Dennett made the legitimate point that the subject who reports experiential phenomena does not have access to the brain mechanisms associated with the occurrence of experiential phenomena; so, how does Dennett know that the 'judgment module' is involved in generating reports about experiential phenomena? Have there been scientific discoveries along these lines; or, is it all just armchair reflection/speculation?

In any case, Strawson rejects the relevance of claims about the judgment module.

> To say that its apparently sensory aspects (say) are in some sense illusory because they are not the product of sensory mechanisms in the way we suppose, but are somehow generated by our processes of judgment, is just to put forward a surprising hypothesis about part of the *mechanism* of this rich seeming that we call experience or consciousness. It is in no way to put in question its existence or reality. (Strawson, 2010, 51-52)

Dennett later itemizes his denial:

> Philosophers have adopted various names for the things in the beholder (or properties of the beholder) that have been supposed to provide a safe home for the colors and the rest of the properties that have been banished from the 'external' world by the triumphs of physics: 'raw feels', 'sensa', 'phenomenal qualities', 'intrinsic properties of conscious experiences', 'the qualitative content of mental states', and, of course, 'qualia', the term I will use. There are subtle differences in how these terms have been defined, but I'm going to ride roughshod over them. In the previous chapter I seemed to be denying that there are any such properties, and for once what seems so is so. I am denying that there are any such properties. But (here comes that theme again) I agree wholeheartedly that there seem to be qualia. (Dennett, 1991, 372)

Once again, Dennett misses the point: that seeming *is* the quale!

Strawson imagines us asking ourselves '*Could there be no experience or consciousness at all?*'

> … The answer 'No' comes quickly and correctly, as it came to Descartes. What is it to suppose that one might be completely wrong? It is to suppose that although it *seems* to one that there is experience – for this can not be denied – there really isn't any experience. But this is an immediate reductio ad absurdum. For this seeming is already experience. (Strawson, 2010, 51)

Dennett seems to understand that the point at issue is the point Descartes made centuries ago. Dennett simply denies that point. He imagines his foil, Otto the Qualiophile, objecting …

> [Otto] It seems to me that you've denied the existence of the most indubitably real phenomena there are: the real seemings that even Descartes in his Meditations couldn't doubt.
>
> [Dennett] In a sense, you're right; that's what I'm denying exist. (Dennett, 1991, 363)

* * *

It is a sad commentary on the state of philosophy of consciousness that one must drag Descartes out of retirement to contest eliminative materialism; but, here we are.

Denying the existence of the red stripe in one's experience while experiencing an afterimage causes considerable collateral damage: one must deny all qualitative aspects of any experience. While Dennett is apparently comfortable with that, I am not.

If the only other choice is eliminative materialism, I choose dualism.

### §2.2.1.2.4 Anticipating Type-Z Materialism

By type-Z materialism I mean any philosophy of (human) consciousness which assumes or from which follows the conclusion that humans satisfy the definition the philosopher's zombie: is physiologically human but does not experience any experiential phenomena (or any qualia, phenomenal qualities, sensa, raw feels or any of the other items whose existence Dennett (1991, 372) denies).

Now, clearly, Dennett denies that humans have or experience anything that may be used to distinguish humans from zombies; so, by my definition, he would be a type-Z materialist.

Of course, Dennett would no doubt deny being a type-Z materialist; indeed, he is well known for claiming that zombies are 'inconceivably preposterous'. Nevertheless, he also says:

> There is another way to address the possibility of zombies, and in some regards I think it is more satisfying. Are zombies possible? They're not just possible, they're actual. We're all zombies. (Dennett, 1991, 406)

Dennett effectively retracts his speculation almost as fast as he makes it. A footnote to the quoted passage reads "It would be an act of desperate intellectual dishonesty to quote this assertion out of context".

What are we to make of this? it makes no sense that I can see for a philosopher to say "Zombies are conceivable, possible and actual; and, we are they! – but don't quote me on that." In any event, others have been more willing to be quoted.

The term 'type-Z materialism' is analogous to the terms 'type-A materialism' and 'type-B materialism' in that they are distinguishable by their responses to Chalmers' zombie argument against materialism.

Briefly, a type-A materialist holds that zombies are inconceivable; whereas, a type-B materialist holds that zombies are conceivable but impossible.

Chalmers never considered the possibility that a materialist might grant that zombies are both conceivable and possible because that conjunction of claims was supposed to entail the falsity of materialism. Others have disagreed, Dave Beisecker wrote:

> This paper explores the viability of rejecting a largely unchallenged third premise of the conceivability argument against materialism. Fittingly labeled 'type-Z' (for zombie), this reply essentially grants to the zombie lover, not just the *possibility* of zombies, but also their actuality. We turn out to be the very creatures Chalmers has taken such great pains to conceive and more conventional materialists have tried to wipe off the face of the planet. So consciousness (at least for us) is a wholly material affair. What is conceivable but non-actual are not zombies, but rather 'angelic' beings possessing an acquaintance with supermaterial phenomenal states. (Beisecker, 2010, 28)

Beisecker acknowledges that type-Z materialism has not been widely considered; although, he mentions Dennett's "We're all zombies" remark as a anticipation of type-Z materialism. However, claims that Dennett ends up as a type-A materialist rather than a type-Z materialist; presumably, because Dennett holds that zombies are inconceivable.

In my view, Bisecker's assessment of Dennett is somewhat off the mark. Type-A materialists may hold that zombies are inconceivable and still affirm

> **[Null-Z] I am not a zombie.**

Beisecker would not affirm [Null-Z] and neither would Dennett.

Crucial for understanding the rhetorical strategy of the type-Z materialist and the flaws of that strategy is Beisecker's clarification of 'supermaterial'.

> 'Super-materialist' is my preferred term for those who follow Chalmers and accept that the zombie argument (or the knowledge argument) demonstrates that there must be more under heaven and earth than is countenanced by (mundane) materialism, for the moniker nicely captures an ambivalence in how the position is understood. To their materialist opponents, supermaterialism advocates us to accept the existence of spooky, supernaturalistic, non-material features of the world, while supermaterialists themselves think they are simply urging us to acknowledge an underlying, intrinsic facet of our material existence, which has heretofore eluded systematic scientific investigation. (Beisecker, 2010, 30, footnote 1)

Surely, this characterization of the dispute between type-Z materialists and their opponents, while seriously overblown, is based on an insight of some significance.

Beisecker is quite right to suggest that Chalmers and other opponents of materialism advocate accepting the existence of something whose existence the materialist must deny; and, that any such a position constitutes dualism; but, he is wrong to suggest that any and every such position counts as substance dualism (the obvious connotation of 'supernatural and non-material').

Chalmers and the earlier Jackson urge us to accept the existence of qualia; but, I

do not read them as saying that a typical quale, for example, the color of an afterimage, is itself a spooky, immaterial, supernatural feature of the world.

Surely there is a less inflammatory reading of anti-materialism, one that preserves Beisecker core claim that the anti-materialistic position is dualistic without exaggerating the degree of dualism necessary to constitute anti-materialism.

*What, then, is the minimally dualistic anti-materialism?*

Beginning in §2.2.2, I will define and defend a position, phenomenon dualism, that is far milder than the specter conjured up by Beisecker. I will show that one need only assert that the color of an afterimage is an experiential phenomenon not identical to any physical phenomenon to constitute an anti-materialistic or anti-physicalistic philosophy.

### §2.2.1.3 Standing Precisely Against Eliminative Materialism

Having rejected eliminative materialism (and type-Z materialism) for the reasons given, I am moved to better define the ground on which I stand before presenting an affirmative argument for the claim that dualism of some sort follows from denying eliminative materialism.

> **[Thesis of Phenomenology] There is First-Person Phenomenology.**

Contrary to Dennett, I hold that there is first-person phenomenology; and, I will assume that this claim is considered uncontroversially true among those who reject eliminative materialism. However, a controversy will emerge once we try to classify the items that first-person phenomenology consists of.

If we take an inventory, we'd find that many commonly discussed items are among the items that an explanation of first-person phenomenology must explain. For example, the philosopher's favorite, phenomenal redness and its cousins; the phenomenal aspects of hearing, tasting, smelling and touching; the raw feel of pain and other bodily sensations – all these and more would be included in an inventory of first-person phenomenology.

For reasons that remain obscure, it seems that many contemporary philosophers hold what I will call the Thesis of Propertyism:

> **[Thesis of Propertyism] The items of which first-person phenomenology consists are themselves properties.**

Phenomenal redness is most commonly thought of as a phenomenal *property*; less commonly as a qualitative property; but, I deny that phenomenal redness is itself a property of any kind. In my view, phenomenal redness is an experiential *phenomenon*; so, in lieu of the Thesis of Propertyism, I hold what I call the Thesis of Phenomenism:

> **[Thesis of Phenomenism] The items of which first-person phenomenology consists are themselves phenomena.**

## §2.2.2 *The Argument for Dualism from Experience*

As noted earlier, Dennett and I share some common ground; but, the road forks at the existence claim. To avoid eliminative materialism, I affirm that the red stripe in my experience while I am experiencing a flag afterimage exists while I am experiencing it. Consequently, I have affirmed a state of affairs from which follows dualism ... *of some sort*.

*What sort of dualism follows?*

I'll now state my argument for dualism from experience; and, we shall see where it leads.

### [1] I am experiencing; therefore, there is experiencing

Experience is a self-verifying fact. It is on the evidence of experience that I assert that I am experiencing or, more generally, that there is experiencing. This is the datum: e*xperiencing occurs*.

The datum is a primitive fact with which scientists and philosophers of consciousness are concerned. It is primitive because it can be neither denied nor doubted. I can not deny experiencing without thereby experiencing denial. Neither can I doubt experiencing without thereby experiencing doubting.

> … *experience is itself the fundamental given natural fact*. Its existence is provably non-illusory because its seeming to exist (which very few deny) is a sufficient condition of its actually existing. (Strawson, 2010, 326)

There a small number of philosophers who *openly* dispute [1]; I am responding to one, Daniel C. Dennett, right now. But, it would also be prudent to anticipate the possibility of an indirect denial. In this respect, we should consider an observation Searle makes early in The Rediscovery of the Mind as he lists various implausible viewpoints.

> Sixth, another extreme view is that maybe consciousness as we normally think of it – as inner, private, subjective, qualitative phenomena of sentience or awareness – does not exist at all. This view is seldom advanced explicitly. Very few people are willing to come right out and say that consciousness does not exist. But it has recently become common for authors to redefine the notion of consciousness so that it no longer refers to actual conscious states, that is, inner, subjective, qualitative, first-person mental states, but rather to publicly observable third-person phenomena. Such authors pretend to think that consciousness exists, but in fact they end up denying its existence. (Searle, 1992, 7)

In what follows, I will use “experience” to refer to the inner, private, subjective, qualitative phenomena that Searle refers to here as “consciousness”. As a singular mass noun, it is appropriate for designating the seamless whole of experience, the phenomenal field or what Searle terms the *unified conscious field*.

> Furthermore, qualitative subjectivity always comes to us as part of a unified conscious field. At any moment you do not just experience the sound of the music and the taste of the beer, but you have both as part of a single, unified conscious field, a subjective awareness of the total conscious experience. So the feature we are trying to explain is qualitative, unified subjectivity. (Searle, 2013)

Usually, in discussing the philosophical implications of subjective experience, we single out from the phenomenal field some item, some piece or *aspect* of experience for further commentary. These bits or aspects of experience are phenomena. I will use *experiential phenomena* to refer to discrete aspects of experience rather than *qualitative phenomena*, the term Searle uses in the passage quoted above. I will use *physical phenomena* for publicly observable, third-person phenomena.

In the case at hand, suppose I induced a flag afterimage while viewing Dennett's video. I see the flag. I see the white wall against which I see the afterimage. I see any number of other things in my field of vision. I hear my dog barking in the background. I smell the pine scent of an air freshener. All of this and more is my experience; but, for purposes of discussion, I focus my attention on one aspect of my experience, one specific red stripe in my experience while I am experiencing a flag afterimage. That red stripe is the aspect of my experience whose existence is at issue in Dennett's example.

*Could it be that experience exists but that none of the aspects of my experience exist?* No. The aspects of my experience, the experiential phenomena that I experience, *are* my experience. If I deny their existence I have denied the existence of my experience.

> The problem of consciousness is identical to the problem of qualia because conscious states are qualitative states right down to the ground. Take away the qualia and there is nothing there. ... Conscious states by definition are inner, qualitative, subjective states of awareness or sentience. (Searle, 1998, 26)

In the case of the red stripe in my experience while I'm experiencing a flag afterimage, the redness and the stripeness are aspects of my experience; so, if their existence is denied, the existence of experience has been denied. Consequently, it follows that

> **[2] An aspect of my experience exists while I am experiencing it.**

An alternate way of defending [2] is to recognize that aspects of my experience are experiential *phenomena* rather than objects or properties, phenomenal, qualitative or otherwise. As an experiential phenomenon, the red stripe in my experience while I am experiencing a flag after image is instantiated by its occurrence to me, its experiencer; hence ...

> **[3] The red stripe in my experience while I am experiencing a flag afterimage exists while I am experiencing it.**

More generally: an experiential phenomenon only exists while being experienced by an experiencing I. Searle's expression of this perspective is well-known.

> Subjective states have a first-person ontology ("ontology" here means mode of existence) because they exist only when they are experienced ... by some "I" that has the experience (Searle, 1998, 41).

The color of an afterimage only exists while being experienced. The redness I see when I am looking at a tomato ceases to exist when I close my eyes; although, the tomato itself continues to reflect light.

It follows that the brain/experience relation is a relation between two distinct

referents each of which exists; but, *which exist in different ways*. An experiential phenomenon exists in an experiencer dependent way. Brain activity is third person, publicly observable or physical phenomena; and, therefore, exists in an experiencer independent way. Something that exists in an experiencer dependent way can not be identical to something that exists in an experiencer independent way.

> **[4] Nothing in my brain is identical to the red stripe in my experience while I am experiencing a flag afterimage.**

A moment's reflection should tell us that there are innumerable pairs of statements analogous to [3] and [4]; for example,

> **[3a] The red stripe in my experience while I am looking at a physically instantiated American flag exists while I am experiencing it.**
>
> **[4a] Nothing in my brain is identical to the red stripe in my experience while I am looking at a physically instantiated American flag.**

An argument for dualism from experience could proceed just as easily from [3a] and [4a] as from [3] and [4]. In each case, we draw the specific conclusion that the experiential phenomenon in question exists while being experienced but is not identical to the physical phenomena, the brain activity, with which it is associated. Generalizing, we may conclude that an experiential phenomenon is not identical to any physical phenomenon in the brain.

This claim is easily extended to physical phenomena existing outside the nervous system. Any physical phenomenon (light reflecting off a tomato, say) that affects my experience does so because it has a causal impact on my nervous system. An experiential phenomenon resulting from such an impact can not be identical to any physical phenomenon that was one of its own causal ancestors.

> **[5] An experiential phenomenon is not identical to any physical phenomenon**

It follows that experiential phenomena and physical phenomena are mutually exclusive categories. In symbolic form (where P represents 'physical phenomenon' and Q represents 'experiential phenomenon' and '≠' means 'is not identical to'):

> **[5.1] $(x)(y)[(Px \ \& \ Qy) \rightarrow (x \neq y)]$**

According to Searle, the failure of the identity relation implies that an experiential phenomenon is not ontologically reducible to a physical phenomenon.

> The basic intuition that underlies the concept of reductionism seem to be the idea that certain things might be shown to be nothing but certain other sorts of things. Reductionism, then, leads to a peculiar form of the identity relation that we might as well call the "nothing-but" relation: in general, A's can be reduced to B's, if A's are nothing but B's. (Searle, 1992, 112)

Intuitively, it makes no sense to say that an experiential phenomenon is nothing

more than something else to which it is not identical; hence, ...

### [6] Experiential phenomena are not ontologically reducible to physical phenomena

Clearly, the next question is whether ontological irreducibility is sufficient for dualism.

Intuitively, if we have two distinct kinds of phenomena, physical and experiential, phenomenon *dualism* seems a more appropriate term for this state of affairs than phenomenon *monism;* so, I answer in the affirmative.

Terminology aside, there is a powerful reason for considering this state of affairs to be dualistic. As any phenomenon is, an experiential phenomenon is *instantiated* by its occurrence; it *exists* while it is occurring. But, it is not identical to *any* physical phenomenon; so, in my view, it counts as a phenomenal individual.

William G. Lycan would consider any assumption that there are phenomenal individuals to be question begging because “contemporary materialists eliminate mental individuals and reduce only states and events” (1987, 17). However, I don't simply *assume* that eliminativism as to experiential phenomena is false, I argued my case for the *conclusions* that Dennett's defense of the eliminative position as to experiential phenomena involved an obvious self-contradiction; and, that anyone who admitted that experiential phenomena occurred would have to admit that experiential phenomena exist while they occur or explain how an experiential phenomenon could possibly occur without existing while it is occurring – a truly bizarre proposition.

Lycan's own views are anti-eliminativist about sensations.

> The 1960s saw heated discussion of Eliminative Materialism in regard to sensations and their phenomenal features. Thus directed, Eliminative Materialism is materialism or physicalism plus the distinctive and truly radical thesis that there have never occurred any sensations; no one has ever experienced a sensation. This view attracted few adherents(!), though to this day some philosophers are Eliminativists with respect to various alleged phenomenal features of sensations. (Lycan, 2014?)

In my argument for dualism from experience, the non-identity claim derives from the existence claims. To recap, something that exists in an experiencer dependent way can't be identical to something that exists in an experiencer independent way. They have different modes of existence.

It may be open to some philosophers to argue that I have not yet shown that an experiential phenomenon would count as a phenomenal individual; but, it isn't clear whether this option is open to Lycan.

> "Qualia" as originally conceived are properties of phenomenal individuals. On this conception, the exemplification of a quale requires at least the ostensible existence of a phenomenal individual (a paradigm of such an individual's bearing a quale would be an after-image's having the vivid, homogeneous color it does). Of course, no materialist admits the existence of phenomenal individuals. (Lycan, 1987, 83)

This is not exactly my position. I don't regard the afterimage as a property bearer; indeed, I deny that I'm talking about properties at all. In my view, an

experiential phenomenon such as the greenish afterimage that Mary sees after staring too long at her tomato, is itself the phenomenal individual in question.

So, the question turns on whether an experiential phenomenon *qua* phenomenon – considered independently of any theory attempting to explain how the phenomenon occurs – can be a phenomenal individual. The stakes are high.

> If there (really) are phenomenal individuals such as sense-data, then materialism is false *right there*; no further reasoning is needed. On the other hand, one is stuck with making a case for phenomenal individuals and with turning aside all of the powerful objections to sense-datum theories. (Lycan, 1987, 18)

I'm not arguing that the color of an afterimage is a sense datum in the sense attributed to Russell, a mental object whose (presumably mental) properties are invoked to explain the phenomenon in question; so, it would be open to Lycan to argue that an experiential phenomenon that is not a sense datum is not a phenomenal individual; or, that, if it is, it doesn't pose a threat to materialism.

I may have lowered the bar for being a phenomenal individual; but, my immediate concern is whether I've achieved dualism; so, I'll defer consideration of the threat my perspective poses for materialism.

> **[7] The ontological irreducibility of experiential phenomena to physical phenomena is sufficient for dualism (of some sort).**

In the end, it may be a judgment call made after counting. If experiential phenomena are not ontologically reducible to physical phenomena, there are two irreducibly distinct kinds of phenomena to be explained; and, intuitively, that suggests phenomenon dualism rather than phenomenon monism.

Nevertheless, however prevalent the conjunction of the existence and the non-identity claims may be, it goes without saying that some philosophers who affirm both claims deny that they are dualists. Such philosophers explicitly or implicitly reject [7]. John R Searle, for example, would affirm [5] and [6] but deny [7]. Later, in §3.2.2, I will examine the arguments he offers to defend himself from allegations of dualism. I will argue that his arguments fail; hence, the denial of [7] is unjustified.

Meanwhile, it remains to be determined what sort of dualism has been derived.

## *§2.2.3 What Sort of Dualism is This?*

I am left with the difficult task of classifying the form of dualism that I have affirmed. I will take my cue from Howard Robinson (2011) who, writing for the Stanford Encyclopedia of Philosophy, SEP, notes that “There are various ways of dividing up kinds of dualism. One natural way is in terms of what sorts of *things* one chooses to be dualistic about”.

I am dualistic about kinds of phenomena. I've spoken about two ontologically irreducible, mutually exclusive, essentially different kinds of phenomena, physical and experiential; and, I've described the brain/experience relation in those terms; so, it seems reasonable to call this perspective ... *phenomenon* dualism.

**[Definition: Phenomenon Dualism] A theory constitutes phenomenon dualism if and only if it *describes* the brain/experience relation using two distinct kinds of phenomena.**

Clearly, in affirming [7] I am affirming that there are two distinct kinds of phenomena. Equally clearly, I am describing the brain/experience relation as the non-identity relation holding between experiential phenomena and the physical phenomena with which experiential phenomena are associated. Given this definition of phenomenon dualism, [7] may be restated as

**[8] The ontological irreducibility of experiential phenomena to physical phenomena is sufficient for phenomenon dualism.**

Robinson does not use this term; so, to avoid the claim that I've merely renamed some well-known form of dualism, I will distinguish it from the forms he mentions and from other well known forms.

### §2.2.3.1 Phenomenon Dualism is Not Predicate Dualism

According to Robinson,

> Predicate dualism is the theory that psychological or mentalistic predicates are (a) essential for a full description of the world and (b) are not reducible to physicalistic predicates. (H. Robinson, 2011)

With *predicate* dualism, the predicates are not reducible; meaning that they are not interchangeable without alteration of meaning. The phenomena themselves *are* reducible.

> Although the predicate ‘hurricane’ is not equivalent to any single description using the language of physics, we believe that each individual hurricane is nothing but a collection of physical atoms behaving in a certain way ... There is token identity between each individual hurricane and a mass of atoms ... (H. Robinson, 2011)

We can take a hurricane as an instance of a weather phenomenon and we can take the swirling mass of atoms that constitute that hurricane as an instance of a physical phenomenon. We then have two sets of phenomena; but, for each instance of that weather phenomenon known as a hurricane, there is an instance of a physical phenomenon (a collection of atoms behaving in a certain way) to which it is token identical; so, the two sets or kinds are not mutually exclusive; and, most importantly, the weather phenomenon is ontologically reducible to the physical phenomenon.

As noted above, there can be no token identity between experiential phenomena and physical phenomena because they have different modes of existence; hence, they are *not* ontologically reducible. It follows that phenomenon dualism makes a stronger claim than predicate dualism. Consequently, while I agree that our language has predicates suitable for referencing physical phenomena and predicates suitable for referencing experiential phenomena and that one set of predicates can not be replaced by the other, I deny being *merely* a predicate dualist.

Phenomenon dualism easily explains why we have and why we need to keep both kinds of predicates: there are two distinct kinds of phenomena to talk about.

### §2.2.3.1.1 Note on Classifying Searlean Philosophy

John Searle has never, to my knowledge, claimed to be a predicate dualist; but, in the midst of defending himself against allegations of property dualism, it sometimes sounds as if he is pleading guilty to a lesser included dualism; so, one argument in support of the intuition that Searle is a dualist of some sort is simply that he goes beyond the limitations of predicate dualism.

> The fact that one and the same conscious state has different levels of description, a level of description where we describe it in terms of its subjective properties, and another level of description where we describe it in terms of its chemical and electromagnetic properties should be no more mysterious to us than the fact that we describe the behavior of a car engine at different levels. The chief difference between the two cases, of course, is that the mental event has a level of description where it is ontologically subjective and that is not the case with the explosion in the cylinder of the car engine. (Searle, 2007a, 176)

Searle is clearly admitting that we have two vocabularies, one for referring to "subjective properties" (experiential phenomena) and the other for referring to electromagnetic properties (physical phenomena); and, presumably, he'd readily agree that we need both vocabularies.

However, because he affirms the ontological irreducibility of experiential phenomena to physical phenomena, he violates the constraint on predicate dualism. Ontological irreducibility entails the denial of token identity; and, that denial is what separates Searle from predicate dualists.

As I will later argue, I support Searle's claim that he is not a property dualist; but, if he is making claims too strong to be considered predicate dualism but too weak to be considered property dualism, the conclusion seems inescapable: there is another form of dualism to be discovered in the *Goldilocks Zone* – not too strong, not too weak, just right.

That form of dualism is *phenomenon dualism*.

### §2.2.3.1.2 Phenomenon Dualism is Not Conceptual Dualism

### §2.2.3.2 Phenomenon Dualism is Not Property Dualism

So far, I have *described* the brain/experience relation as a relation between experiential phenomena and the physical phenomena with which they are correlated but to which they are not identical. Is there a definition of 'property dualism' such that an affirmation like the one I've just made – one devoid of any mention of any properties of any kind – is sufficient to constitute *property* ***dualism***? I would like to think that the answer is obvious; but, as we shall see, the situation is murky.

According to Robinson:

> … property dualism says that there are two essentially different kinds of property out in the world.

As a first approximation, this seems reasonable enough; but, it has a weakness: It does not explicitly require that the two kinds of property be used in an explanatory role.

Imposing such a constraint on the role played by the pair of property types stems from wanting to preserve the traditional structure of an explanation.

> Properties are typically introduced to help *explain* or *account for* phenomena of philosophical interest, especially in doing ontology. The existence of properties, we are told, would explain qualitative recurrence or help account for our ability to agree about the instances of general terms like 'red.' In the terminologies of bygone eras, properties save the phenomena; they afford a *fundamentum in re* for things like the applicability of general terms. Nowadays philosophers make a similar point when they argue that some phenomenon holds *because of* or *in virtue of* this or that property, that a property is its *foundation* or *ground*, or that a property is the *truthmaker* for a sentence about it. These expressions signify explanations. (Swoyer and Orilia, 2011)

If we add an assumption that properties are not free-floating but are always properties of some substance or object (the property bearer), we have the traditional structure of an explanation: properties are attributed to substances/objects to explain phenomena.

> **[Definition: Property Dualism] A theory constitutes property dualism if and only if it *explains* the brain/experience relation using two distinct sets of properties, where one set is the set of properties used by scientists to explain physical phenomena.**

By requiring two *sets* of properties, I hope to avoid disputes based on arbitrarily defined *kinds*. Without this constraint one could, under Robinson's definition, classify all physicists as property dualists merely because they have discovered that subatomic particles have two essentially distinct kinds of properties, static and dynamic.

My definition of substance dualism builds on the definition of property dualism.

> **[Definition: Substance Dualism] A theory constitutes substance dualism if and only if (1) it *explains* the brain/experience relation using two (or more) distinct sets of properties, where one set is the set of properties used by scientists to explain physical phenomena; and, (2) it assumes or concludes that two distinct kinds of property bearers are required to bear all the properties required by the theory.**

As I have defined my terms, phenomenon dualism is not sufficient to constitute property dualism because phenomenon dualism merely describes the brain/experience relation. It states what is to be explained; but, does not itself specify how many sets of properties or how many kinds of property bearers will be required to explain all instances of each kind of phenomena.

Naturally, once an explanation is offered, it is possible that the resulting theory might constitute property or substance dualism; but, no one is required to offer their own explanation of experiential phenomena; and, I do not. I prefer to follow Searle's advice concerning the division of labor between scientists and philosophers: "… where a scientific solution is at least possible, the philosophical

task is to prepare the problem conceptually, to get it into a kind of shape where it admits of being treated as a scientific problem" (Searle, 2007a, 169).

Scientists may eventually produce an explanation of experiential phenomena; but, there is no guarantee that they will succeed. If they do not, philosophers will have to choose between living with a profound mystery and advancing their own explanations.[5]

Naturally, some philosophers whom I would classify as phenomenon dualists might not want to wait for scientists to either claim success or admit failure. Chalmers, for example, admits that there are two kinds of phenomena to be explained; and, proposes an explanation for that state of affairs that involves postulating properties unknown to scientists. Such a theory certainly constitutes both property dualism and phenomenon dualism.

Another philosopher might propose an explanation that did not involve properties unknown to scientists. Searle, for example, holds that experiential phenomena will eventually be explained as being due to the biological properties of the brain. Such a theory certainly constitutes phenomenon dualism but does not constitute property dualism.

The fact that phenomenon dualists who choose to offer their own explanations for experiential phenomena have their choice between explanations that are property monistic and those that are property dualistic means that phenomenon dualism does not entail and is not sufficient to constitute property dualism.

Since my definition of substance dualism builds on my definition of property dualism, if phenomenon dualism does not constitute or entail property dualism it will not constitute or entail substance dualism either.

> **[9] Phenomenon Dualism does not constitute and does not entail either property dualism or substance dualism.**

Someone might contest [9] simply by assuming a different set of definitions. In particular, someone might assume a definition of property dualism that includes both phenomenon dualism and property dualism as I define them.

I believe this to be a mistake because it suppresses the ability to distinguish a form of dualism that may turn out to be compatible with science as we know it from a form of dualism that expands the ontology of science.

The practical effect of distinguishing what is commonly conflated is that we could justify placing Searle and Chalmers in different camps. Both recognize dualism in the explanandum; but, only Chalmers questions the ability of

---

5 The philosopher of consciousness may have a role even after the science of consciousness reaches its limits. Kozuch and Kriegel (forthcoming) argue that "the identification of a neural feature that correlated perfectly with consciousness would still leave open a certain metaphysical question: is the relation between consciousness and the relevant neural feature merely correlation, or is that correlation indicative of a deeper, more intimate relation between the two? Work addressing this further question can be thought of as attempting a philosophical interpretation of scientific theories, somewhat on a par, say, with philosophical interpretations of quantum mechanics: in both cases, philosophy has to take over where science proper ends in order to articulate an intelligible conception of how the world must be given what the science suggests.".

scientists to explain experiential phenomenon without expanding the ontology of science to include a new type of property, phenomenal or protophenomenal, that is unknown to scientists.

The presence of phenomenon dualism in contrast to reductive/eliminative forms of materialism or physicalism, turns on the contents of the explanandum. Is experiential phenomena not identical to physical phenomena among the items to be explained by a proposed theory of the brain/experience relation? If so, there is phenomenon dualism. If not, no one is trying to explain the relevant phenomena.

My definition of property dualism turns on the contents of the explanans, the means by which an explanation is achieved – one set of properties or two. By specifying one set of properties as the set of properties used by scientists to explain physical phenomena, I've linked the definition to a claim of explanatory power made by materialists and physicalists. For example, Dennett writes:

> According to the materialists, we can (in principle!) account for every mental phenomenon using the same physical principles, laws and raw materials that suffice to explain radioactivity, continental drift, photosynthesis, reproduction, nutrition and growth. (Dennett, 1991, 33)

I have three objections to Dennett's version of the explanatory power claim.

1. The category of mental phenomena is broader than the category of experiential phenomena;
2. Dennett denies the existence of experiential phenomena; and,
3. There is an ambiguity as to the referent of “we”. Does he mean that we, humans, can explain mental phenomena because human scientists can explain mental phenomena or because human philosophers could explain mental phenomena to their own satisfaction (if scientists fail to do so)?

My reformulated claim of explanatory power focuses on experiential phenomena precisely because these are the phenomena that present the greatest challenge to materialism. It also assumes the existence of experiential phenomena not identical to physical phenomena. Not everyone accepts that; but, without this assumption, no one is even attempting to address the relevant issue.

I've also removed the reference to a particular philosophy, materialism; and, finally, the claim of explanatory power is no longer about philosophy. It's now a claim about the Explanatory Power of Science, EPS.

> **[EPS] Scientists can explain or account for the existence of experiential phenomena not identical to physical phenomena using only the objects, properties, principles, laws and raw materials with which scientists explain physical phenomena.**

By itself, phenomenon dualism makes no claim as to *whether* or *how* the existence of experiential phenomena not identical to physical phenomena will be explained.

Optimistic mysterians assume that there is a risk that scientists will fail. Both property dualists and pessimistic mysterians assume that scientists will not succeed. Property dualists go on to offer explanations that postulate properties

unknown to scientists.

Given a suitably inclusive definition of “physicalism”, one might claim to be a physicalist on the grounds that experiential phenomena will, some sweet day in the future, be explained by physical scientists. Should that ever occur, the philosophical questions would remain: Has phenomenon dualism been embraced or denied? Has phenomenon dualism been explained or explained away?

#### §2.2.3.4 Phenomenon Dualism or an Atypical Event Dualism?

Phenomenon dualism is easily translated into an event dualism, provided that

1. An event is understood as the occurrence of a phenomenon; and,
2. A phenomenon is instantiated by its occurrence.

However, in currently versions of event dualism, an event is understood as the instantiation of a property at a time (Kim, 1976); and, this is not always a merely linguistic difference. There are differences between *what* I am saying and what typical event dualists are saying. These differences are related to how we define an event; and, would survive translating phenomenon dualism into an atypical event dualism.

##### §2.2.3.4.1 Garret on Event Dualism

Garrett (2000, 393) writes that

1. Event dualists hold that mental events are not physical events; and,
2. Nonphysical events are events instantiating nonphysical properties.

The first claim seems reasonable enough; but, I reject the second.

For me, the occurrence of an experiential phenomenon is an experiential event and the occurrence of an physical phenomenon is a physical event. Assuming that the non-identity of physical and experiential phenomena translates into the non-identity of physical and experiential events, phenomenon dualism could be viewed as an event dualism. But, it bears repeating in this context that, in asserting phenomenon dualism, I make no claims about the nature of the properties that may be invoked to explain the occurrence of an experiential phenomenon. It may be that all the properties in virtue of which I experience phenomenal redness or some other experiential phenomenon are physical properties.

##### §2.2.3.4.2 Robinson's Dualism of Neural and Qualitative Events

William S. Robinson is hoping to show that dualism is a viable option and to make it more palatable. I have a lot of sympathy for his efforts; and, it's possible that, by elaborating my view, I too will help make dualism more palatable than eliminative materialism or identity theory physicalism.

> The dualism to be discussed in this paper claims that there are phenomenal qualities, and that these are different from, and not composable from, the properties and relations found in our natural sciences. (W. Robinson, 2014, 156)

I can agree with that; but, ...

> The distinctive claim is that our experiences (a) are non-physical, qualitative events; that is, they are, or essentially involve, instances of phenomenal qualities, i.e. instances of properties that are not instances of physical properties; and (b) these property instances have no further, or hidden, physical nature. (W. Robinson, 2014, 157)

Although “qualitative event” sounds like it could be used as a synonym for “experiential event”, I deny the identity claim that Robinson makes. For me, a phenomenal quality is an experiential *phenomenon* not a *property* of any kind.

In other respects, my views are analogous to Robinson's. In my view, my acquaintance with experiential phenomena reveals their essential nature. This is the thesis of revelation. However, since I assume that in experiencing I am acquainted with phenomena rather than properties, it is easy to assume that there is some physical phenomenon correlated with an experiential phenomenon in which I am interested. They may even be so closely tied together that the experiential phenomenon is the appearance to me of the underlying physical phenomenon; but, an appearance is not identical to the reality of which it is merely an appearance.

Nevertheless, I would still object to the claim that a certain neural firing pattern in the C-fibers of the brain is the hidden physical nature of the experiential phenomenon known as pain. For one thing, it's misleading to say that discharging C-Fibers are themselves hidden in any useful sense; any technician with sufficient training and the right technology can find discharging C-Fibers.

That said, a claim about the “hidden nature” of experiential phenomena has significance when read as a claim that the true nature of the connection between correlated experiential and physical phenomena is largely unknown (and possibly unknowable); although, both Robinson and I would say that we already know that the relation is not identity.

#### §2.2.3.4.3 Robinson's Epiphenomenalism

Both Robinson and I are qualia realists, as are identity theory physicalists. We all agree that qualia are real. Robinson parts company with identity theory physicalists by denying physical/phenomenal identity (Robinson, 2012), which the phenomenon dualist also denies.

Robinson disagree as to whether a quale is a property or a phenomenon. Robinson assumes that a quale is a property, a qualitative property or phenomenal property. This creates a minor linguistic difficulty for Robinson when he adds what he calls the signature claim of epiphenomenalism, that “experiences do not cause anything” (Robinson, 2012, 147). The difficulty is that claiming that a quale is qualitative property which is epiphenomenal implies that something can be an epiphenomenon without being a phenomenon.

In my view, only a phenomenon can be an epiphenomenon; but, it is easy to read Robinson as saying that a quale is a (qualitative) property which is causally inert.

In my version of phenomenon dualism, experiential phenomena play a causal role; albeit, a minor one in most cases. However, the definition of phenomenon

dualism neither requires or prohibits a claim of qualia epiphenomenalism. So, but for the category error (classifying a quale as a property rather than a phenomenon), Robinson would be a phenomenon dualist who affirms qualia epiphenomenalism; whereas, I am phenomenon dualist who denies qualia epiphenomenalism.

#### §2.2.3.4.4 Atypical Event Dualism

These comparisons to previously known forms of event dualism should suffice to distinguish them from the event dualism into which phenomenon dualism could be translated; so, phenomenon dualism could only be considered a new or atypical form of event dualism.

In any case, I prefer the term phenomenon dualism over the term event dualism for the following reasons:

1. Phenomenon dualism as I've developed it was inspired by the philosophy of John R. Searle who contrasted physical and mental phenomena rather than physical and mental events; and,
2. My argument for dualism from experience concerns the ontological status of experiential phenomena rather than the events constituted by their occurrence.

The second point bears some elaboration. If someone were to argue that a physical event and an experiential event were identical, we would need to examine the identity conditions for an event in order to evaluate that claim. Suppose we say that an event is the occurrence of a phenomenon at a time. Now, suppose we find that at time t, a physical phenomenon, C-fibers firing, and an experiential phenomenon, are co-instantiated.

Instead of being two distinct events, could they be one self-identical psychophysical event? Sure, if one defined a psychophysical event as an event that instantiated a physical phenomenon and an experiential phenomenon, you could claim to have psychophysical event monism. But, such a definition would not make the experiential phenomenon identical to the physical phenomenon; so, the taxonomic question would remain. Is it dualism, monism, a dual-aspect theory or something else?

### §2.2.3.6 Conclusion

Phenomenon dualism is not just another, more familiar form of dualism renamed. It fills a needed gap in the taxonomy of dualisms; something more serious than predicate or conceptual dualism but not as serious as property dualism.

Of all the forms of dualism considered, phenomenon dualism is most like an event dualism with a non-standard definition of an event. If an event can be defined as the occurrence of a phenomenon, phenomenon dualism can be translated into an event dualism and back again. Indeed, phenomenon dualism and the event dualism into which it is translated would be complementary perspectives.

* * *

I will now consider objections to the Argument for Dualism from Experience, ADE.

## §3 Objections to the Argument For Dualism From Experience

Ontological reducibility is a strong claim; so, making ontological reducibility the criteria for achieving physicalism makes physicalism difficult to achieve. Identity theory physicalism would be the only form of physicalism; but, (identity theory) physicalism would be a sufficient defense to allegations of dualism. Philosophers who wish to be considered physicalists despite not being identity theory physicalists may make physicalism easier to achieve by weakening their notion of reducibility so that an ontological reduction is not necessary for non-identity theory physicalism or what I will call *weak physicalism*. One question that arises from these efforts is whether a weak physicalism provides a sufficient defense to allegations of dualism. This question will be considered in §3.1.

Ontological irreducibility is a weak claim, requiring only the non-identity of physical and experiential phenomena. And, it's an easy claim to make; provided, you aren't bothered by the unpopularity of dualism. All you have to do is affirm what Dennett denies: that an experiential phenomenon exists while it is occurring to its experiencer. To make dualism harder to achieve and easier to avoid, someone who affirms [5] and [6] but wants to deny [7] might require arbitrarily more than ontological irreducibility for achieving dualism. This option will be considered in §3.2.

Finally, in §3.3, I'll consider direct challenges to [5] by considering arguments an identity theorist might make in defending materialism/physicalism from advocates of the knowledge argument.

## §3.1 Is Physicalism a Sufficient Defense against Dualism?

Physicalists seem to accept some version of an argument that may be formulated as follows:

1. Whatever is is physical.
2. Consciousness is.
3. (therefore) Consciousness is physical.

The antithesis to this argument is that not all that is is in the same way; so, *even if all that is is physical, not all that is physical is physical in the same way*.

Suppose that we considered brain activity to be physical because the brain exists in an experiencer independent way and brain activity is measurable in an experiencer independent way. Suppose further that we considered experiential phenomena to be physical because, despite existing in an experiencer dependent way, they are caused by brain activity.

Given that this perspective constitutes a form of physicalism, the question is whether it also constitutes phenomenon monism or phenomenon dualism.

* * *

Philosophers who admit the ontological irreducibility of experiential phenomena and who want to call themselves physicalists may do so simply by adopting a

form of *non-identity physicalism*; meaning, a form of physicalism other than identity theory physicalism.

For non-identity physicalists, any of a number of claims may serve as the basis of the claim of physicalism. For example, it may be claimed that experiential phenomena are caused, constituted or realized by physical phenomena; or, that experiential phenomena supervene on or emerge from physical phenomena.

From the point of view of someone reviewing the claims of non-identity physicalists who deny also being phenomenon dualists, the relevant question is whether the basis for the claim of physicalism is sufficient to defend the non-identity physicalist against allegations of dualism.

### *§3.1.1 Is Weakly Reductive Physicalism a Defense?*

Some philosophers believe that physicalism must be reductive.

> A physicalist view of the mind must be reductive in one or both of the following senses: it must identify mental phenomena with physical phenomena (ontological reduction) or it must give an explanation of mental phenomena in physical terms (explanatory or conceptual reduction). (Crane, 2000, 73)

I will use *strong reduction* as a synonym for ontological reduction where that requires making an identity claim; and, *weak reduction* for any form of reduction that may be alleged without asserting an identity claim. I will use *strongly reductive physicalism* as a synonym for identity theory physicalism; and, *weakly reductive physicalism* as a term for any form of non-identity physicalism that purports to be reductive without being strongly reductive.

I will assume that a strongly reductive physicalism is a sufficient defense against allegations of dualism stronger than conceptual dualism; and, for present purposes, I will assume that a weakly reductive physicalism is a sufficient defense against allegations of dualism stronger than phenomenon dualism. After all, if scientists do someday manage to give an explanation of experiential phenomena in physical terms, there will be no need to postulate non-physical properties or non-physical substances/objects to explain experiential phenomena; and, claims of property dualism or substance dualism will be claims in search of an alternate motivation.

For those evaluating the claims of weakly reductive physicalists who deny being dualists, the crucial question is: *How does a weakly reductive physicalism provide a sufficient defense to allegations of phenomenon dualism?*

For those evaluating the claims of the phenomenon dualism, the crucial question is: *How does phenomenon dualism avoid ceasing to be dualism when found to be consistent with a weak physicalism?*

I'll answer on behalf of the phenomenon dualist in the hope of clarifying the process of evaluating answers to these questions.

Let us start our consideration with a special case, the dispute between identity theory physicalism and phenomenon dualism. Everything turns on the claim of

physical/phenomenal identity. If the identity theorist shows that for each experiential phenomenon there is a physical phenomenon to which it is identical, strongly reductive physicalism has been achieved and phenomenon dualism has been refuted.

On the other hand, if the claim of physical/phenomenal identity fails, strongly reductive physicalism is refuted; and, we have two fundamentally distinct kinds of phenomena, a state of affairs that I take to constitute phenomenon dualism.

The physicalist can now lower the standard for achieving physicalism so that an ontological reduction is no longer required, offering instead the possibility that scientists may someday explain the occurrence of experiential phenomena.

I don't object to lowering the standard for achieving physicalism; but, I reject the implicit assumption that, when such is done, the standard for achieving dualism is automatically raised so that achieving physicalism (by whatever means) avoids phenomenon dualism (or any more serious form of dualism).

In my view, the test for physicalism and the for dualism are independent tests; and, that leaves open the possibility that a philosophical position other than identity theory physicalism will pass both tests resulting in a conclusion that a given position is both dualistic and physicalistic. Thus, various forms of non-identity physicalism pass my test for phenomenon dualism, having two fundamentally distinct kinds of phenomena.

It is open to the non-identity physicalist to argue that the test for physicalism and the test for dualism are linked such that passing the test for physicalism precludes also passing the test for dualism.

In the next section, I will consider ways in which physicalists implicitly assume a linkage between testing for physicalism and testing for dualism.

### §3.1.2.1 Scientific and Philosophical Perspectives

Kriegel offers the helpful insight that scientists and philosophers have different perspectives, including different criteria for defining materialism.

> This view – sometimes referred to as *emergentism* – that consciousness is caused by the brain, or causally emerges from brain activity, is often taken by scientists to be materialist enough. But philosophers, being interested in the *ontology* rather than the *genealogy* of consciousness, commonly take it to be a form of dualism. If consciousness cannot be shown to be itself material, but only caused by matter, then consciousness is itself immaterial, as the dualist claims. At the same time, the position implicit in scientists' work is often that what is caused by physical causes in accordance with already known physical laws should be immediately considered physical. (Kriegel, 2007a, 55, fn 12)

Kriegel says that the position he attributes to scientists is “not unreasonable” and, in another work, a review of the last book by the late Jeffrey Gray, goes on to say …

> The position implicit in Gray's book, and probably in most scientists' working conception of their project, may be called *inclusive physicalism*. Unlike reductive physicalism, the view is not that consciousness will eventually turn out to be nothing but some kind of

> brain activity. Rather the idea is that the phenomenon of consciousness will eventually be shown to be fully caused by brain activity in a lawlike manner consistent with the already established laws of neuropsychology (and physics). In that eventuality, the scientist's reasoning implicitly goes, it would be legitimate to consider consciousness a physical phenomenon. The operative principle here is that a phenomenon that can be explained as fully causally determined by physical phenomena, in obedience to physical laws, ought to be considered itself a physical phenomenon. The project, then, is to 'physicalize' consciousness, not reductively however, but 'inclusively', i.e. by showing that a kinder, more inclusive conception of the physical is warranted that would treat consciousness as a physical phenomenon. (Kriegel, 2007b, 98-99]

Kriegel's intuitions about the different perspectives of scientists and philosophers seem to be empirically testable. Perhaps experimental philosophers and/or cognitive scientists will someday query scientists and philosophers to assess their attitudes. However, until we have actual data to the contrary, I will assume that most scientists and philosophers would agree that, even if an afterimage is physical in some sense, it is not physical in the same sense that a stone is physical.

> **[Non-Allness] Even if all that is is physical, not all that is physical is physical in the same way.**

The question that arose concerning Searlean dualism is still with us; although, in a slightly altered form: Does having items that are physical in different ways – because they have different modes of existence, one experiencer dependent and the other experiencer independent – constitute dualism?

While it may be reasonable for some purposes to expand the notion of physical in this manner, doing so introduces the possibility of equivocation as to what is physical or material. So, when necessary to facilitate clarity, I will adopt the following conventions for using suffixes to convey a more precise meaning.

I will use numeric subscripts as type indicators to designate what is physical (or material) according to some standard – yielding ways of being physical (or types of physicality, if you prefer).

> $\text{Physical}_1$ =df. Something that is said to be physical because it is exists in a publicly accessible, experiencer independent way.
>
> $\text{Physical}_2$ =df. Something that is not $\text{physical}_1$ because it exists in an experiencer dependent way but which is said to be physical for some other reason.

I will use alphabetic subscripts to denote the cumulative content of the “physical world” in terms of physical types designated with numeric suffixes. Thus,

> $\text{Physical}_A$ =df. The world consisting of that which is $\text{physical}_1$ and nothing more.
>
> $\text{Physical}_B$ =df. The world consisting that which is $\text{physical}_1$ plus that which is $\text{physical}_2$.

Now, in the exclusive or restrictive sense, “physical” refers only to items that are $\text{physical}_1$ (anything in the $\text{physical}_A$ world). In the inclusive or expansive sense, “physical” can refer to anything that is either $\text{physical}_1$ or $\text{physical}_2$ (anything in

the physical$_B$ world).[6]

This notation may make it easier to see the flaw in Searle's attempt to deny being a property dualist on the grounds that he denies the mutual exclusivity of physical and experiential (mental, in his terms) phenomena.

The non-identity of experiential phenomena and physical$_1$ phenomena makes those categories mutually exclusive and is the basis for the conclusion of ontological irreducibility of experience. Searle accepts that claim of ontological irreducibility. To avoid the conclusion of dualism that seems to follow, Searle shifts to denying mutual exclusivity by saying that experiential (physical$_2$) phenomena are included within the world of physical$_B$ phenomena.

Without that equivocation, Searle is in agreement with dualists as to the crucial point that experiential phenomena are not physical$_1$ phenomena. They are within the physical$_B$ world but not the physical$_A$ world. Further, Searle disagrees with the materialist for whom the physical$_A$ world is all there is.

Setting aside questions as to which philosophical arguments presuppose this sort of equivocation, one may wonder why we don't just define dualism to require postulating something that isn't physical in any sense. The answer is that we have not yet exhausted all the ways of being physical; so, we may need to extend the notation still further.

If physicists were to postulate an immaterial consciousness to explain the laws of physics (as some say they already have), a theory of experience that invoked both the physical brain and that immaterial consciousness to explain experiential phenomena would be compatible with physics and could be considered a form of physicalism. We would need additional rules for ascribing the predicate *physical*:

> Physical$_3$ =df. Something that is not physical$_1$ or physical$_2$ but is postulated to help explain the laws of physics pertaining to that which is physical$_1$ and/or physical$_2$.
>
> Physical$_C$ =df. The world consisting of that which is physical$_1$, physical$_2$ and physical$_3$.

While such a theory could be considered physicalist in some sense, it would also be dualistic. Indeed, such a theory would fit the traditional concept of interactive substance dualism.

Nevertheless, until there is an empirical falsification of the von Neumann/Wigner interpretation of quantum mechanics (that consciousness causes the collapse of the wave function), there is a risk that we will end up with a theory that includes something physical$_3$. Given a sufficiently expansive conception of “physical”, such a theory would constitute a form of physicalism; but, I reserve the right to call it dualism even as I also reserve the right to call it a form of physicalism.

---

6 Throughout this paper, the unqualified term “physical” defaults to the physical$_1$ meaning when talking about particular items or to the physical$_A$ meaning when talking about the world or the contents of the world.

### §3.1.3.2 Dualism as a Matter of Counting

I readily concede that there are circumstances that cry out for such an expansive use of “physical”. For example, when confronted with the possibility that physicists may turn in a so-called theory of everything that does nothing more that supply an elaborate correlation of the physical and the phenomenal, Chalmers would award them a grade of “Incomplete”. If not all that is happening has been explained, the job's not done; so, it would be helpful in this circumstance to be able to say that not all that is (allegedly) $\text{physical}_2$ has been explained.

Similarly, it would also be helpful to be able to say that the the potential for scientific explanation means that state of affairs I'm calling phenomenon dualism is not necessarily inconsistent with holding a scientific worldview; particularly if one remembers that, if Kriegel is correct, scientists do not limit themselves to explaining experiential phenomena by stating identity claims that would support ontological reductions.

Nevertheless, the question of whether some form of dualism holds is a matter of counting how many kinds of relevant items there are.

If scientists ultimately succeed in explaining experiential phenomena on the basis of whatever is $\text{physical}_1$, we could say that scientists have explained how physical causes have phenomenal effects; and, therefore, that scientists have explained phenomenon dualism.

The alternative, would be to define dualism with respect to what scientists have or may someday explain. From this perspective, it could be said that experiential phenomena are “physical” *now* because scientists *may someday discover* that they are $\text{physical}_2$ – by showing how experiential phenomena are caused or generated by $\text{physical}_1$ phenomena.

The implicature of this perspective is that, if experiential phenomena are someday shown to be $\text{physical}_2$ phenomena, the appearance of dualism will have been explained away. But, if scientists show that experiential phenomena are caused by $\text{physical}_1$ phenomena, only the need to invoke $\text{non-physical}_1$ properties or $\text{non-physical}_1$ substances to explain the occurrence of $\text{physical}_2$ (experiential) phenomena will disappear.

The appearance of dualism would certainly remain. The dispute as to whether having two distinct kinds of phenomena constitutes phenomenon dualism or phenomenon monism would likely continue.

### §3.1.3.3 Dualistic Physicalism is not Self-Contradictory

Clearly, the premise of weak physicalism is that a claim of $\text{physical}_1$/phenomenal identity is not necessary for physicalism; but, the issue at hand is whether a claim of $\text{physical}_1$/phenomenal non-identity is *sufficient for dualism*.

In addition, it seems that weak physicalists implicitly assume that a claim of physicalism is sufficient for non-dualism. However, we are not dealing with an analytic truth. One isn't able to inspect the meaning of the word “physical” and determine that it means “not dualism” the way that one can inspect the meaning

of the word “bachelor” and determine that it means “an unmarried male”.

Consequently, one may, without self-contradiction, hold *both* that ontological reducibility (a claim of physical$_1$/phenomenal identity) is not necessary for physicalism; *and*, that ontological irreducibility (a claim of physical$_1$/phenomenal non-identity) is sufficient for dualism.

Holding both theses would constitute a dualistic physicalism.

## *§3.1.2 Constitutional Physicalism*

> In earlier physicalist literature, the ‘is’ in the phrase ‘the mental is physical’ was understood as the ‘is’ of strict identity. But recently physicalists have tended to understand the ‘is’ as something closer to the ‘is’ of constitution. To say that everything is physical in this sense is to say that everything either is a physical entity or is constituted by or composed of physical entities. (Crane, 1995, 212)

Constitutionalists often explain the nature of the relation of material constitution by considering the relation between a statue and the material of which it is made.

Consider the puzzle posed by the actions of Artemis, the master goldsmith of Avalon. On Monday, Artemis buys a lump of gold which he names *Lump*. On Tuesday, he forms it into a statute of Pegasus which he names *Pegasus*. Constitutionalists argue that the statue is not identical to the lump of gold from which it is made. If, come Wednesday, Artemis melts the statue down into an ingot, *Ingot*, the statute would be destroyed but the gold would endure. To the constitutionalist, it seems reasonable to say that the statue and the gold have different modal properties. Hence, by Leibniz's Law (Indiscernibility of Identicals), it follows that the statue is not identical to the gold of which it is made.

This is quite a dilemma. If the lump of gold and the statue are not numerically identical, the constitutionalist is committed to saying that there are two objects where other philosophers (myself included) can only see one. If Pegasus is identical to the gold of which it is made, then so is Lump and Ingot. But then, by the transitivity of identity, Pegasus, Lump and Ingot would each be identical to the other two, which is absurd.

The way out of this dilemma is to affirm that Pegasus is not identical to the gold of which it is made; but, that it *is* identical to the object constituted when that gold is cast into the form of that statue. If we assume that a material object is constituted by matter *and form*, the problem of modal properties disappears.

Come Wednesday, Artemis melts down Pegasus. Pegasus is destroyed. The gold remains as it did before; but, the object constituted by the gold and the form of the statue is also destroyed. The statue and the object that is that-gold-in-that-form came into existence at the same time and are destroyed at the same time.

Now the constitutionalist may still want to say that the Pegasus is constituted by the gold and the form into which it is fashioned. If so, I would agree. But, I would go on to say that Pegasus is identical to the material object constituted by the gold of which it is made and the form into which that gold is put.

It is difficult to see how the constitutionalist could deny identity statements of this sort. Pegasus is a material object; otherwise, it wouldn't have a material constitution. But then there is a material object to which Pegasus is identical; namely, itself.

Thus, it would appear that constitution is identity after all; although, one must articulate a theory of material objects before this becomes apparent; and, one must be content with token identities.

The conclusion is that constitutionalists are making *an* identity claim; although, not the simplistic claim that a constituted object is identical to its material components alone. This conclusion is bolstered claims that suggest that constitutionalists are aiming for an ontological reduction.

> … constitution does not amount to identity, so when A constitutes B, A and B are numerically distinct, and we have here as many entities as we do when A causes B. On the other hand, on a natural understanding of the notion of constitution, when A constitutes B we are justified in saying that B is "nothing but" A … (Kozuch and Kriegel, 2015, 413)

How can a B be nothing but something else, an A to which it is not identical?

If it is true that constitutionalists are making an identity claim, it is doubtful that constitution theory has anything to contribute to our understanding of the brain/experience relation that isn't being said by identity theory.

Consider the afterimage.

The phenomenon dualist agrees with Dennett that an afterimage does not exist as a material object; but, rejects the eliminative approach by saying that it does exist while it is occurring to its experiencer.

*What does the constitutionalist say?*

Constitutionalists might be eliminative as to experiential phenomena not identical to physical$_1$ phenomena. Alternately, they might hold that an afterimage is a material object in some state or undergoing some process; but, that would be to affirm rather than deny physical/phenomenal identity theory.

I am not aware of any constitution theorist who crosses the line the phenomenon dualist steps across: affirming the existence of the afterimage while denying that it is self-identical to *any* physical$_1$ phenomenon (or property or object, for that matter).

## §3.2 Ontological Irreducibility is Insufficient for Dualism

It seems intuitively obvious (to me, anyway) that Searle is a dualist of some sort. For one thing, he articulates Cartesian phenomenology better than anyone else who denies being a dualist; so, if there is anyone who deserves to be considered as a dualist who denies being a dualist, it is Searle.

Nevertheless, I am puzzled by the long-standing dispute between Searle and his critics as to whether he is a *property* dualist. The accusation is out of proportion to the evidence. Searle asserts the existence of two kinds of *phenomena*. His critics accuse him of *property* dualism.

In this section, I will examine the case for Searlean dualism and many of Searle's defenses, including defenses which seemingly argue that something more than ontological irreducibility is required for dualism. I will argue that, while the ontological irreducibility of experiential *phenomena* is sufficient for phenomenon dualism; and, that Searle's defenses are not adequate to defend himself against that charge.

I will, however, defer consideration of Searle's attempt to argue that his thesis of causal reducibility protects him from allegations of dualism. It doesn't seem possible to discuss the causal reducibility of experiential phenomena without discussing the causal reducibility of the experiencing subject; so, I'll return to the thesis of causal reducibility after shifting to focus on the brain/subject relation.

## §3.2.1 The Case for Searlean Dualism

Searle is more than a predicate dualist because he denies the identity of experiential (first person) phenomena and physical (third person) phenomena; but, in my view, he has an ironclad defense to charges of property dualism: he holds that one set of properties is sufficient to explain both kinds of phenomena. “All of our mental states are caused by neurobiological processes in the brain” (Searle, 2006, 40).

To avoid a gap in the taxonomy of dualisms, there must be a form of dualism that is stronger than predicate/conceptual dualism but not as strong as property dualism. I call it *phenomenon dualism*.

### §3.2.1.1 The Evidence of Phenomenon Dualism

Searle is well-known for asserting that there two kinds of phenomena.

> Where consciousness is concerned, there are first-person phenomena and third-person phenomena. (Searle, 2007a, p. 177)

Searle offers a simple, indeed elegant, reason for holding that the two kinds of phenomena are *essentially* different from each other: they have different modes of existence.

> Conscious states exist only when they are experienced by some human or animal subject. In that sense, they are essentially subjective … Subjective states have a first-person ontology ("ontology" here means mode of existence) because they exist only when they are experienced by some human or animal agent. They are experienced by some "I" that has the experience, and it is in this sense that they have a first person ontology. (Searle, 1998, p. 40-41)

Something that exists in an experiencer dependent way can't be identical to something that exists in an experiencer independent way; so, these must be distinct, mutually exclusive categories. Searle draws the reasonable conclusion that the two kinds of phenomena are irreducibly distinct.

> You can't reduce these first-person subjective experiences to third-person phenomena for the same reason that you can't reduce third-person phenomena to subjective experiences. You can neither reduce the neuron firings to the feelings nor the feelings

> to the neuron firings, because in each case you would leave out the objectivity or the subjectivity that is in question. (Searle, 1997, p. 212)

All of this justifies the conclusion that Searle is a phenomenon dualist as I have defined that term.

### §3.2.1.2 The Leap from Phenomenon Dualism to Property Dualism

Having conceded the non-identity of the physical and the experiential, from which follows the ontologically irreducibility of the two, Searle struggles to avoid the dualism that seems to follow.

Searle's critics accuse him of being a *property* dualist; but, it is not clear how critics justify the leap from the evidence of phenomenon dualism to the accusation of property dualism. For example, Edward Feser sums up his case against Searle, thus:

> If paradigmatically and uncontroversially physical phenomena are essentially objective, and paradigmatically and uncontroversially mental phenomena are irreducibly subjective, then it follows that they are of fundamentally different metaphysical kinds. It follows, that is, that property dualism - the claim that there are (at least) two metaphysically fundamental kinds of property in the universe - is true. (Feser, 2004)

Precisely *how* does it follow? As it stands, Feser's argument is incomplete.

> [F-1] Searle postulates two ontologically distinct kinds of phenomena.
>
> [F-2] ???
>
> [F-3] (therefore) Searle is a property dualist.

[F-1] is certainly true; but, [F-3] does not follow from [F-1] alone. Feser could fill in [F-2] with a sufficiency claim, such as

> [F-2a] Postulating two ontologically distinct types of phenomena is *sufficient* to constitute property dualism.

However, his argument does not support a sufficiency claim.

> The bottom line is that the distinction on which property dualism rests - that between irreducibly subjective and objective phenomena - is one that Searle himself is committed to as marking out two objective categories of phenomena in the universe. (Feser, 2004)

I agree that property dualism rests on the distinction between subjective, first-person phenomena and objective, third-person phenomena; but, that only establishes a *necessary* condition for property dualism. I also agree that Searle is committed to the reality of that distinction; but, that only establishes that Searle is committed to a necessary condition of property dualism.

As I have defined it, property dualism requires postulating two sets of properties to explain occurrences of the two kinds of phenomena. Feser is free to adopt a different definition, of course; but, he should state the definition he's using to classify Searle as a property dualist. Curiously enough, he doesn't.

> If the physical processes which cause consciousness are objective third-person phenomena, and consciousness and other mental phenomena are subjective or first-

> person in nature, it is reasonable to describe the latter as being of a fundamentally different kind than the former. That is, it is reasonable to say that there exists in the universe a dualism of properties. (Feser, 2004)

Once again, I agree that it is reasonable to say that first-person phenomena and third-person phenomena are of fundamentally different kinds; and, that Searle does say that. But, how do we get from there to a dualism of *properties*? Feser doesn't say.

#### §3.2.1.3 Searle is Not a Property Dualist

That a critic such as Feser would accuse Searle of property dualism without offering a formal definition of property dualism is strange. Stranger still is that Searle has neither demanded that his critics state the definition by which they say he is a property dualist nor faulted them for having failed to do so. Strangest of all is that Searle has never stated his own definition of property dualism and then argued that his views don't fit that definition. Instead, Searle distinguishes his views from those of property dualists (or dualists generally) by stating a proposition that he says is true of them but not of him.

Before considering these defensive arguments, I want to make a few brief comments on the case for Searlean property dualism.

It is clear that the evidence that Feser cites to support the claim of property dualism is the same evidence that I cite to support classifying Searle as a phenomenon dualist; but, while we agree that there is evidence of dualism, I do not consider these to be alternate names for the same state of affairs.

Recognizing phenomenon dualism as a category distinct from property dualism has two advantages. First, it allows us to put Searle and Chalmers in different camps based on the number of sets of properties required to explain both kinds of phenomena. Second, it makes available an additional argument against the identity theory, the argument for an appearance/reality distinction between experiential phenomena and any physical phenomena with which they may be correlated.

### §3.2.2 Evaluating Searle's Defenses

Searle is effectively saying that, despite having acknowledged the ontological irreducibility of experiential phenomena to physical phenomena, he is not a (property) dualist because something more is required for property dualism.

In “Why I Am Not a Property Dualist” Searle offers four propositions which he says are true of property dualists but which he denies are true of him. I'll now consider the first three of these, deferring a consideration of the fourth (which concerns epiphenomenalism) until the sections on the brain/subject relation.

#### §3.2.2.1 The Denial of Mutual Exclusivity

In the first proposition, Searle attributes to property dualists the belief that physical phenomena and experiential phenomena are mutually exclusive categories.

> [PD-1] There are two mutually exclusive metaphysical categories that constitute all of empirical reality: they are physical phenomena and mental phenomena. Physical phenomena are essentially objective in the sense that they exist apart from any subjective experiences of humans or animals. Mental phenomena are subjective, in the sense that they exist only as experienced by human or animal agents. (Searle, 2002a, p. 58)

Searle has two reasons for distinguishing his views from an affirmation of [PD-1].

First, he denies [PD-1] because such things as “declines in interest rates, points scored in football games, reasons for being suspicious of quantified modal logic, and election results in Florida” are not easily classified as exclusively mental or physical. Searle seems to be assuming that the categories of physical phenomena and mental phenomena must be both mutually exclusive and collectively exhaustive to constitute property dualism.

Arguably, a claim of ontological irreducibility would support a conclusion of mutual exclusivity. Searle is suggesting that something more is required to make the categories collectively exhaustive.

However, if we limit our consideration to phenomena *relevant to the task at hand*, discerning the nature of the brain/experience relation, Searle's views *would* satisfy his own requirement that the categories of physical phenomena (which exists in an experiencer independent way) and experiential phenomena (which exists in an experiencer dependent way) be collectively exhaustive as well as mutually exclusive. I can't think of a category of phenomena relevant to discovering the nature of the brain experience relation that exists without existing in either an experiencer independent way or an experiencer dependent way but not both.

Second, Searle argues that there is only one world but that there are any number of ways of dividing up the world – all of them interest relative.

> There are not two (or five or seven) fundamental ontological categories, rather the act of categorization itself is always interest relative. ... We live in exactly one world and there are as many different ways of dividing it as you like. (Searle, 2002a, p. 59)

Let us assume arguendo that the structure of our categories is interest relative; and, that our interest at the moment is in discerning the true nature of the brain/experience relation. How many categories would we need to use to talk intelligently about this interest of ours? We would need two, the same two that Searle uses.

> ... the problem with all forms of materialism is they are confronted with the question: Are there two kinds of phenomena going on in there [or] only one? And the answer has to be: Where consciousness is concerned, there are first-person phenomena and third-person phenomena. Materialism is forced to say there is only one kind of thing, the third person phenomena. (Searle, 2007a, p. 177)

Here, Searle asks the crucial question. And, the answer he gives, whether interest relative or not, is the correct answer: two.

Consequently, I can't distinguish Searle's position from an affirmation of [PD-1] together with a self-contradictory denial of mutual exclusivity, its key claim.

Hence, I conclude that this defense to allegations of dualism fails.

However, just because I think that Searle's defense fails, it doesn't follow that I think that the allegation of *property* dualism succeeds. In my view, the allegations of property dualism fail for another reason; namely, that Searle only recognizes one kind of property, physical (although some physical properties are also biological). Nevertheless, even if cleared of the charge of property dualism, Searle may still be guilty of the lesser included dualism, phenomenon dualism.

The basis for allegations of Searlean dualism is Searle's claim that there are two kinds of *phenomena*; and, I can see no plausible reason for holding that such a position constitutes phenomenon monism rather than phenomenon dualism.

### §3.2.2.2 Distinctness Arguments

The second of the propositions that are supposed to discriminate between property dualists and Searle's own view is

> [PD-2] Because mental states are not reducible to neurobiological states, they are something *distinct from* and *over and above* neurobiological states. The irreducibility of the mental to the physical, of consciousness to neurobiology, is by itself sufficient proof of the distinctness of the mental, and proof that the mental is something over and above the neurobiological. (Searle, 2002a, p. 59)

Unfortunately, Searle doesn't define “distinct from”. I take that phrase to mean “not identical to”. So, if an experiential phenomenon is not identical to any physical phenomenon it is by definition distinct from every physical phenomenon. Searle is free to use a different definition of “distinct from”; but, if he does so, he should specify the definition according to which experiential phenomena are not distinct from the physical phenomena to which they are not identical.

Similarly, it is not clear why asserting the non-identity of consciousness and the brain is not sufficient for asserting that consciousness is something over and above its neurological substrate.

According to Searle's own analysis of concepts of reduction, an ontological reduction occurs when something A is shown to be nothing but something B; and, he criticizes identity theorists for their reductive intentions.

> … historically the identity theorists that I know, with very few exceptions, had a reductionist motive. They wanted to get rid of subjectivity. They wanted to say that consciousness is *nothing but* neurobiological states of the brain neurobiologically described in third-person terms. I have argued in this article that we know independently that that claim is false. (Searle, 2007a, p. 177)

Clearly, then, Searle would say that it is not the case that consciousness is nothing but a neurobiological state of the brain; therefore, there is no ontological reduction. But, that is no different from saying that consciousness is something over and above the neurobiological state of the brain.

Further, Searle undermines his attempt to show that he holds that consciousness is nothing but a neurobiological state. After noting that there are two kinds of phenomena, first-person and third-person, Searle writes

> Materialism is forced to say there is only one kind of thing, the third person phenomena, but we all *know* from our experiences, *that in addition to the neuron firings*, the computer programs, the behavior, etc., *there are my subjective, qualitative conscious states*. (Searle, 2007a, p. 177 (emphasis supplied))

If subjective, qualitative, conscious states are something in addition to the neuron firings, it is hard to understand how consciousness qua experience is not something over and above those neuron firings. It is much easier to conclude that Searle is contradicting his earlier claims by affirming the content of [PD-2].

The unanswered question remains: *What more besides non-identity is required for dualism?*

Searle (2007a, 175) tells us that dualists postulate distinct domains for the mental and the physical; but, he does not define *domain* or *domain distinctness*; so, it's not clear why the non-identity that is the basis for ontological irreducibility is not sufficient for "the distinctness of the mental" which, presumably, *is* sufficient for dualism.

Searle is free to choose a definition of "domain" such that domain distinctness does not follow from the ontological irreducibility to which Searle is already committed, thereby making domain distinctness an additional requirement for dualism. However, Searle would then have to explain away what appears to be a self-contradiction. In arguing that a science of consciousness is possible, he writes "You can have a perfectly objective science of an ontologically subjective domain." (Searle, 2007a, p. 175). Logically, the domain of the ontologically subjective would have to be distinct from the domain of the ontologically objective; they contain items with different modes of existence.

In my view, each kind of phenomenon is a domain; and, domain distinctness is an additional conclusion that follows from the non-identity of experiential and physical phenomena, the basis for the conclusion of ontological irreducibility to which Searle and I are both committed.

We are at an impasse, slowly sinking into a quagmire, trying to conduct a philosophical inquiry that turns on a number of undefined terms. The alternative is to avoid being sucked into the quagmire in the first place, by recognizing that the non-identity of physical and experiential phenomena is sufficient for ontological irreducibility which is sufficient for dualism.

### §3.2.2.3 The Real Objection is Interaction

> It is now time to try to say what exactly is wrong with dualism. I have already said that consciousness is not ontologically reducible to brain processes. Isn't that already a kind of dualism? (Searle, 2007a, 175).

Searle now tells us that "The real objection to dualism is that we cannot give a coherent account of reality on dualist assumptions." (Searle, 2007a, 175). By this he means that, if we put brain processes and consciousness in different domains, it becomes difficult to explain their interactions.

> … it becomes difficult, if not impossible, to explain how brain processes in one ontological domain could cause consciousness in another ontological domain. … it is

> hard, if not impossible, to see how consciousness could have any causal impact in the world (Searle, 2007a, 175).

In my view, Searle is quite correct about the difficulty of explaining the interaction of brain and consciousness; but, that difficulty is there from the moment one affirms the non-identity claim that is the basis for the claim of ontological irreducibility.

Once we grant that there are two kinds of phenomena, we generate the problem of explaining their interactions; but, the difficulty of explaining their interactions is unaffected by any meta-theoretical or taxonomic dispute as to whether holding that there are two distinct kinds of phenomena constitutes phenomenon dualism or phenomenon monism (or biological naturalism, non-reductive materialism or whatever).

Consequently, the difficulty of explaining brain/consciousness interactions is not a reason for denying [7], that the ontological irreducibility of experiential phenomena to physical phenomena is sufficient for dualism (of some sort).

### §3.2.2.4 Contesting the Language of Discourse

The third of the propositions that are supposed to discriminate between property dualists and Searle's own view is

> [PD-3] Mental phenomena do not constitute separate objects or substances, but rather are features or properties of the composite entity, which is a human being or an animal. So any conscious animal, such as a human being, will have two sorts of properties, mental properties and physical properties.

In his attempt to show that this proposition is true of the property dualist but not true of him, Searle argues that the traditional language of discourse “was designed to contrast the mental and the physical” (2002, 61) and that makes it difficult or impossible for him to say what he wants to say without appearing to contradict himself.

Specifically, there is

> … a false presupposition in the very terminology in which we stated the problem. The terminology of mental and physical, of materialism and dualism, of spirit and flesh, contains a false presupposition that these must name mutually exclusive categories of reality – that our conscious states qua subjective, private, qualitative, etc. cannot be ordinary physical, biological features of our brain. (Searle, 2006, 39-40)

Now, it certainly *appears* that Searle is contradicting himself when he complains about the assumption that the mental and the physical are distinct. As noted above when making the case for Searlean dualism, Searle provides an elegant way to show that the physical and the experiential are mutually exclusive categories: they have different modes of existence.

*Is Searle's diagnosis sound?*

Is there some flaw in the traditional language of discourse that makes it look like Searle contradicts himself when he does not; or, is there some virtue in the traditional language of discourse that enables us to notice when Searle contradicts himself? I'm inclined to the latter answer because I can't imagine

how something that exists in an experiencer independent way could be nothing other than something that exists in an experiencer dependent way.

That said, a detailed review of Searle's argument against the language of discourse is worthwhile because *it fails instructively*. There *is* a flaw in the traditional language of discourse; but, it is one that can be surgically removed without killing the patient.

### §3.2.2.4.1 The Intuition of Distinctness

The traditional language of discourse does, indeed, contrast the mental and the physical just as Searle says; but, this was not the arbitrary result of a coin toss back at Plato's Academy and continued since that time by nothing better than inertia. The traditional language reflects what staunch physicalist David Papineau calls the intuition of distinctness.

> I would say that it strikes most people as obvious that the conscious mind is something more than the brain and that physicalism is therefore false. (Papineau, 2008, 57)

> ... there is something very counter-intuitive about the phenomenal-material identity claims advocated by materialists. When materialists urge that *seeing red* (and here you must imagine the redness) is identical to some material *brain property*, it strikes many people that this *must* be wrong. From now on I shall call this natural reaction the 'intuition of mind-brain distinctness'. (Papineau, 2002, 74)

Papineau and any number of other physicalists respond by attempting to provide Wittgensteinian therapy for or to otherwise explain away the intuition of distinctness; but, Searle provides an elegant antidote to all such efforts. At least for the subset of the mental known as the experiential, the intuition of distinctness is true. What exists with a subjective mode of existence can't be identical to what exists with an objective mode of existence.

So, the complaint about the language of discourse fails if it remains as initially stated, as a complaint about the contrast between the mental or experiential and the physical. Searle, however, spends more time trying to overcome the contrast between materialism and dualism; and, here, I think he has a point.

According to Searle (2002, 62-63; 2007a, 178), materialists and dualists each say something true and something false; whereas, his own perspective, biological naturalism, preserves the truths offered by both materialists and dualists while rejecting their falsehoods. Specifically, Searle claims that

1. Materialists claim (truly) that the universe consists entirely of material phenomena such as physical particles in fields of force; ***and***, (falsely) that ontologically irreducible states of consciousness do not exist; whereas,
2. Dualists claim (truly) that ontologically irreducible states of consciousness exist; ***and***, (falsely) that they are not ordinary parts of the physical world.
3. Biological Naturalists claim (1) that the universe consists entirely of material phenomena such as physical particles in fields of force; ***and***, (2) that ontologically irreducible states of consciousness exist; ***and***, (3) that they are ordinary parts of the physical world.

Two questions naturally arise at this point:

1. Does Searle actually endorse the truths advocated by both materialism and dualism without endorsing their falsehoods?
2. If so, does this fact have the effect Searle claims for it, that of making Searle's position *neither* materialistic nor dualistic (rather than *both* materialistic and dualistic)?

I will consider these questions in § 3.2.2.4.2 and §3.2.2.4.3, respectively.

### §3.2.2.4.2 Equivocation or Self-Contradiction?

It seems that Searle's attempt to distinguish his position from dualism appears to turns on an equivocation in the meaning of the term “physical world” that occurs in the statement of the dualist's position and in the statement of the biological naturalist's position.

*Is the ordinary physical world limited to third person phenomena?*

As noted earlier, Searle argues that materialists are wrong to say that there is only one kind of phenomenon, third-person phenomena; so, in stating the first clause of the biological materialist's position, he contradicts himself by saying that materialists are correct to say that the universe consists entirely of material phenomena.

Similarly, given that experiential phenomena are ontologically irreducible to physical phenomena because they have different modes of existence, both Searle and the dualist are correct to affirm that such phenomena exist; but, the dualist is also correct to say that experiential phenomena are not among the third person, material phenomena making up the ordinary physical world. If Searle disagrees with the dualist at this point, he would be contradicting his claim that ontologically irreducible states of consciousness exist.

*Does the physical world include both third-person and first-person phenomena?*

If so, then both the dualist and Searle would say that the physical world includes first person phenomena as ontologically irreducible states of consciousness; but, Searle would be contradicting his claim to have agreed with materialists that the world consists entirely of material (third-person) phenomena.

In my view, Searle equivocates between a restrictive conception of the physical world as consisting only of third-person phenomena and an expansive conception in which both first-person and third-person phenomena are included as ordinary parts of the physical world. The restrictive conception is necessary to support the conclusion of ontological irreducibility; but, the expansive conception is necessary for Searle's defense to the allegations of dualism that seem to follow from the claim of ontological irreducibility.

Absent this equivocation, Searle can't make good on his claim that he agrees with the truths of materialism and dualism but rejects their falsehoods; so, this defense to allegations of dualism fails.

### §3.2.2.4.3 An Alternate Way of Overcoming the Opposition

Even if Searle can show that biological naturalism rejects the falsehoods of both materialism and dualism while preserving their truths, it doesn't follow that biological naturalism is *neither* a form of materialism nor a form of dualism. Until one rules out the possibility of dualistic materialism, one could argue that biological naturalism is *both* dualistic and materialistic because Searle preserves the truth of each position while rejecting their falsehoods.

Searle began by complaining that the traditional language of discourse evolved to contrast the mental and the physical; but, his linguistic defense to allegations of dualism is based on overcoming the contrast he makes between materialism and dualism.

That's not the same contrast.

I've argued above that Searle himself provides a strong reason for maintaining the traditional contrast between the physical and the experiential: first-person and third-person phenomena have different modes of existence. Nevertheless, I can easily imagine wanting to overcome the contrast between materialism and dualism.

Tim Crane appears to disagree.

> Dualism can be contrasted with monism, and also with physicalism. It is argued here that what is essential to physicalism is not just its denial of dualism, but the epistemological and ontological authority it gives to physical science. (Crane, 2000, 73)

However, Crane made it clear that he was defining dualism as either substance dualism or property dualism; so, he didn't explicitly address the question of whether physicalists must deny lesser forms of dualism as well.

I will be more explicit. I reject the assumption that denying all forms of dualism is essential to physicalism.

A position like the phenomenal concepts strategy (a.k.a. conceptual dualism) is best understood as a form of identity theory physicalism that acknowledges its conceptual dualism to avoid stronger, more objectionable forms of dualism. Consequently, I see no reason to deny that it is a (mildly) dualistic form of physicalism.

Phenomenon dualism, the the state of affairs in which two ontologically distinct kinds of phenomena are invoked to describe the brain/experience relation, makes a claim stronger than the claim made by conceptual dualism. Indeed, one could say that phenomenon dualism explains why we have two sets of terms: there are two kinds of phenomena to talk about.[7]

Phenomenon dualism is not compatible with a *strongly reductive physicalism;*

---

7 In contrast, conceptual dualists hold that both sets of terms refer to brain activity. To anticipate the discussion of the Knowledge Argument, both the phenomenon dualist and the conceptual dualist would admit that we have two sets of terms; for example, *pain* and *firing C-fibers*. The conceptual dualist holds that each term of any such pair refers to the same brain activity as the other. The phenomenon dualism holds that the referents of each of the paired terms are distinct. One is an experiential phenomenon and the other is a physical phenomenon.

meaning, a form of physicalism that asserts that there is only one kind of object, physical objects; that physical objects have only one kind of property, physical properties; and, that there is only one kind of phenomenon to be explained, physical (third-person) phenomena.

Phenomenon dualism denies the identity claim that supports the ontological reduction of experiential (first-person) to physical$_1$ (third-person) phenomena required for strong physicalism. Consequently, a strongly reductive physicalism provides an adequate defense against claims of phenomenon dualism; and, phenomenon dualism is a sufficient refutation of strongly reductive physicalism.

However, phenomenon dualism is not necessarily incompatible with a *weakly reductive physicalism*; meaning, a philosophy of consciousness whose concept of reduction is weaker than that of ontological reduction because it does not require the ontological reduction (identity) of experiential and physical phenomena.

Consider a form of physicalism which holds that scientists will eventually explain experiential phenomena in physical terms without showing that experiential phenomena are ontologically reducible to physical phenomena; that philosophers may consider such an explanation to be an explanatory or conceptual reduction; and, that philosophers may conclude that experiential phenomena so explained are, in some sense, physical phenomena despite not being physical$_1$ phenomena.

Such a philosophy shares with phenomenon dualism the non-identity claim that is the basis for the conclusions of ontological irreducibility and dualism. It also has a very reasonable basis for being considered a form of physicalism.

*What prevents us from saying that such a philosophy is a dualistic physicalism?*

Only the convention that assumes a necessary opposition between physicalism and dualism. This may be the convention; but, there is no logical reason why one must adopt it; and, I decline to do so.

I have no objection to the claim that a weakly reductive theory of consciousness might still constitute a form of physicalism. However, I reject the assumption that any form of physicalism no matter how weakly reductive is as adequate a defense to dualism as a strongly reductive physicalism. Without that assumption, one must apply one's tests for physicalism/materialism and for dualism independently of each other, leaving open the possibility that a philosophy of consciousness could be both physicalistic (or materialistic) and dualistic.

If scientists someday explain experiential phenomena in physical terms, weakly reductive physicalists will have good reasons for feeling vindicated. Their prediction that physical scientists will eventually be able to explain experiential phenomena will have been proven right. But, unless it happens that the scientific explanation for experiential phenomena shows that for every experiential, first-person phenomenon there is a third-person, physical phenomenon to which it is identical, we will continue to have phenomenon dualism rather than phenomenon monism.

Some philosophers have taken advantage of this possibility. For example, Joseph Almog calls his view *dualistic materialism* because it incorporated four theses.

One pair of these theses are

> … driven by dualism and asserted numerical and by-nature distinctions of the mental and the physical. They are complemented by another pair, driven by materialism, insisting on structured mental/physical connections: modal (necessary) and by-nature inter-dependence. (Almog, 2010, 362)

These four theses come from two intuitions: the *duality intuition* and the *structural connection* intuition. The first is easy to pinpoint.

> I am a common sense dualist in thinking my pain and my firing of C fibers make a duo of numerically distinct kinds of phenomena. Whether they make two substances or kinds of properties is a further, more theoretical issue, of logical grammar. But in whatever level of the type hierarchy we end up, they make *two* items of that level. (Almog, 2010, 355)

The structural connection intuition is harder to pin down; but, it involves saying that the material realm is the only realm of being there is; and, that there is a lawful connection between the material and the phenomenal. Almog denies the assumption that zombie arguments make, that fixing the physical facts does not fix the phenomenal facts; or, as some put it, that after fixing the physical facts God had to do something else to fix the phenomenal facts.

> Not so, when the physics of the cosmos attains a certain complexity and is embedded in a certain niche, *it* – any further acts of God aside – is the engenderer of the emergent phenomena. (Almog, 2010, 362)

Interestingly, these intuitions must be kept in a dynamic balance like yin and yang. Too much emphasis on the duality intuition leads to an extreme form of dualism, substance dualism. Too much emphasis on the structural connection intuition leads to an extreme form of materialism, identity theory.

The parallels between Almog's dualistic materialism and Searle's biological naturalism are instructive. Both have the same position with respect to the intuition of materialism. Searle holds that it is a mistake to think that consciousness is something added on to physical reality.

> On my view, given the constitution of reality, consciousness has to follow in the same way that any other biological property, such as mitosis, meiosis, photosynthesis, digestion, lactation, or the secretion of bile, follows. (Searle, 2007a, 177)

Searle himself is not likely to agree that biological naturalism is a form of dualistic materialism; nevertheless, one may accept that there is some truth in Searle's complaints about the language of discourse without accepting Searle's remedy for the problematic opposition between materialism and dualism. There is an alternate remedy: instead of abandoning the traditional language of discourse in its entirety, simply abandon the assumption that no philosophy of consciousness can be both materialistic (or physicalistic) and dualistic.

### §3.2.2.5 Conclusion

Searle is well known for admitting the ontological irreducibility of consciousness; and, for defending himself from allegations of dualism based on that admission. I've examined several of Searle's defenses and I've argued that they fail to show

that Searle is not a dualist; hence, they fail to show that ontological irreducibility is insufficient for dualism.

The last of Searle's defenses to be considered is the thesis of causal reducibility; but, I'll defer consideration of it until I reach the section on the brain/subject relation since it concerns Searle's effort to defend subject causation without being accused of dualism.

## §3.3 Brain/Experience Identity Is A Viable Theory

It might be objected that I have simply assumed the truth of the non-identity claim that Dennett built into flag afterimage scenario; and, that I need to defend the non-identity claim against challenges to [5] posed by identity theorists.

Well, I will consider challenges to [5]; but, my affirmative defense of the claim of physical$_1$/phenomenal non-identity is that I have not simply assumed it. In my view it follows from the existence claims. Items which exist with different modes of existence can't be identical.

### *§3.3.1 Theory vs. Meta-Theory*

The claim that humans experience phenomena which are not identical to any physical phenomenon is a philosophical theory about humans, as is its negation. A philosophical theory which aims to classify other philosophical theories based on the claims those theories make about humans is a *meta-theory*.

In the case at hand, philosophers may agree as to the crucial conjunction of the existence claim and the non-identity claim but disagree as to the classification of their position – as physicalism, dualistic physicalism, non-reductive physicalism, subjective physicalism, phenomenon dualism or whatever.

Obviously, a debate over whether dualism follows from [5] emphasizes a natural philosophical concern with the meta-theoretical (taxonomic) classification of philosophical positions; but, in my view, the relative merits of identity and non-identity theories of the brain/experience relation should depend on what these theories say about humans rather than what they say about themselves.

Consequently, to facilitate that shift in rhetorical emphasis, I will situate the discussion in the context of debates concerning the Knowledge Argument, KA, where the focus is on what may be said about Mary.

In what follows, I will first present a short history of Jackson's KA highlighting the Jackson/Churchland dialogue that led to a clarification of ambiguities in the original presentation. Following that is a brief discussion of the need for a further clarification – as to what constitutes success. The KA is supposed to be an argument against physicalism; but, with so many versions of physicalism available, it's reasonable to assume that the KA might succeed against some targets but not all.

Next I will present a weakened version of the KA that targets *phenomenon monism*, those forms of physicalism that assume that there is only one kind of phenomenon. Finally, I will argue that the revised KA achieves its purpose,

justifying the rejection of phenomenon monism, by showing that there is some truth (some fact, some information or some knowledge) that (1) survives the story told by eliminativists; and, (2) escapes the story told by identity theorists.

## *§3.3.2 The Knowledge Argument*

The KA had a long history before Frank Jackson came along with a version that fired the imagination of philosophers in a way that no previous version had ever done. See Stoljar and Nagasawa (2004) for a review this history. I will only consider the KA from Jackson on.

### §3.3.2.1 Jackson's Knowledge Argument

Frank Jackson's knowledge argument poses a challenge for physicalism. Mary has been raised her entire life in a colorless room connected to the outside world only by black and white television monitors (Jackson, 1982). To update the gedanken, we might imagine that she also has complete access to the Internet; but, that she only has a black and white computer monitor.

Mary grows up to become a brilliant scientist specializing in color vision and, we are asked to assume, knows the physical facts about color vision; meaning, all the facts that can be taught by lessons or learned by her own independently conducted scientific research. Other ways of characterizing the knowledge she has include "knowledge by description", "knowing-that", "propositional knowledge" and "discursive knowledge".

Eventually, Mary is released from her confinement and she experiences seeing colors for the first time. Upon seeing a ripe tomato we can imagine her exclaiming, "Ah! So, ***that*** is tomato red" or "***This*** is what it is like to see red" or something similar. We can easily imagine Mary tweeting the Friends of Mary network, "I am now experiencing tomato red" (where “tomato red” is her name for that shade of phenomenal redness she experiences when looking at a ripe tomato.)

After staring at the tomato in total fascination for about a minute, Mary looks at a white wall and experiences her first color-complement afterimage. From her studies she knows that the complement of red is green; but, she doesn't know the name for the particular shade of green she is presented with. However, being the clever scientist that she is, she came prepared for this little experiment. She emerged from her room with a plan to name the shade of green the afterimage presents to her *otamot green* – spelling “tomato” backwards to signify color complementation.

> 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, 1982, 130)

Physicalism is falsified because “qualia are left out of the physicalist story”.

### §3.3.2.2 The Charge of Equivocation

Paul Churchland (1985) pointed out each of Jackson's two premises relies on a different way of knowing; and, he identified the two ways of knowing that the argument draws on as knowledge by description (the way that Mary knows the physical facts) and knowledge by acquaintance (the way that Mary knows experiential phenomena).

Churchland then argues that, if the premises are revised to eliminate equivocation between the two ways of knowing, the conclusion of the KA does not follow from the revised premises because they are insufficient to rule out the possibility of an identity between what is known by description and what is known by acquaintance.

I agree with Churchland – up to a point. After controlling for equivocation between the ways of knowing, the conclusion does not follow from the two premises alone; but, that doesn't break the KA. It merely means that advocates of the KA will have to work for their conclusions, possibly by introducing further assumptions, clarifications, definitions or whatever. Opponents of the KA will have to work for any conclusions they try to reach; so, I don't see it as a problem that proponents of the KA will have to work for their conclusions.

That said, it must also be said that opponents of the KA have taken some damage from the Churchland defense; their argumentative burden has been increased by the concessions Churchland makes.

> The materialist can freely admit that one has "knowledge" of one's sensations in a way that is independent of the scientific theories one has learned. This does not mean that sensations are beyond the reach of physical science. (Churchland, 1985, 24)

The way of knowing that Churchland refers to in this passage is knowing by acquaintance; so, having admitted there is knowing by acquaintance, the physicalist must now give a account of knowing by acquaintance consistent with physicalism. Furthermore, it now becomes difficult if not impossible to argue that Mary learns nothing at all after first becoming acquainted with phenomenal redness.

Churchland is certainly correct as to the epistemological point he makes. Merely admitting that knowing by acquaintance is a feature or a capacity of the knowing/experiencing subject does not by itself entail that “sensations are beyond the reach of physical science” (whatever that phrase may mean). But, physicalists would have the burden of showing that experiential phenomena such as sensations have been brought within the reach of physical science in a way that doesn't threaten their preferred version of physicalism.

Identity theory physicalists would have an additional burden well, that of justifying the inference that the relation between an experiential phenomenon and a physical phenomenon is the identity relation rather than a non-identity relation such as appearance/reality. By the terms of the gedanken, Mary has all the physical facts. We can imagine her assembling her facts in a vast Catalog of Neurology and Phenomenology, CNP, in which each distinctive neural firing pattern is listed and numbered for easy reference. Let us suppose that, beside each entry, is a mention of any experiential phenomena, associated with that

neural firing pattern. For senses other than vision, Mary could have filled in much of this information from her own experience; but, for vision she relied on the reports of subjects who reported what they saw while Mary recorded the output of her neuroscopic readings.

In becoming acquainted with tomato red, Mary becomes acquainted with the experiential phenomenon associated with a particular previously cataloged neural firing pattern, NFP-4738, say. How does Mary justify the conclusion that the experiential phenomenon only she experiences is identical to the neural firing pattern in her brain that anyone could measure with the right instruments?

### §3.3.2.3 The Total Knowledge Response

In "What Mary Didn't Know", Jackson (1986) replied to Churchland and other critics. He rejected the reformulation of KA that Churchland (1985) had proposed and provided his own.

> [KA-1] Mary (before her release) knows everything physical there is to know about other people.
>
> [KA-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).

When she experiences phenomenal redness, Mary experiences something new to her. Learning by experiencing (knowing by acquaintance) gives Mary a new fact about the experiences of others. Since she had all the physical facts relating to color vision, if she acquires a fact she didn't have before, it can not be a physical fact about the experiences of others.

> Therefore,
>
> [KA-3] There are truths about other people (and herself) which escape the physicalist story. (Jackson, 1986, 293)

Significantly, Jackson admitted that each of the premises relies on a different way of knowing; and, that Churchland had correctly identified them: knowledge by description in [KA-1] and knowledge by acquaintance in [KA-2]. However, Jackson denied the charge of equivocation by saying that the KA is about Mary's total knowledge however acquired. With respect to [KA-2], Jackson claims that

> ... Mary learns something on her release, she acquires knowledge, and that entails that her knowledge beforehand (what she knew, never mind whether by description, acquaintance, or whatever) was incomplete. (Jackson, 1986, 294)

This reply answers the charge of equivocating between two ways of knowing; but, it does not clearly address Churchland's other point. If two (or more) ways of knowing are involved in acquiring Mary's knowledge, one can't simply *assume* that the knowledge gained by acquaintance consists of facts that *escape the physicalist story*, the conclusion to which the knowledge argument would lead us. On the other hand, until we hear the tale as told by the physicalist, we can't assume that physicalism accounts for phenomenal knowledge or even that it attempts to do so.

### §3.3.2.4 Further Clarification and Concessions

Churchland (1989) acknowledged that each of the premises of the reformulated knowledge argument is plausibly true; provided, that knowing was understood as knowing by description in the first premise but as knowing by acquaintance in the second.

Churchland again argues that, after controlling for equivocation, the conclusion doesn't follow from the premises; but, he also recognizes that the debate isn't over and takes a significant step toward identifying the issue on which the outcome turns: “whether sensory qualia form a metaphysically distinct class of ***phenomena*** beyond the scope of physical science”. (Churchland, 1989, 74 (*my emphasis*))

This is a significant concession to those friends of the KA who discuss it in terms of phenomena rather than properties. Once the items under discussion, the items of which first-person phenomenology consists, are brought under the concept of *phenomenon* rather than *property,* we recover the use of the appearance/reality distinction.

Opponents of the KA also suffer other setbacks from Churchland's efforts to tame it.

> If it works at all, Jackson's argument works against physicalism not because of some defect that is unique to physicalism; *it works because no amount of discursive knowledge, on any topic, will constitute the nondiscursive form of knowledge that Mary lacks*. (Churchland, 1989, 72)

I read this passage as an admission that discursive knowledge about the brain, knowledge by description, won't generate an instance of actual experiencing, an instance of knowledge by acquaintance. No matter how extensive her discursive knowledge is, Mary will not know what it is like to experience tomato red until she becomes acquainted with that experiential phenomenon … by actually experiencing it.

In one sense this should be fairly obvious, as can be seen by considering the way in which humans, as a species, are like Mary before her release. Humans are trichromats; we have three kinds of cones in the retina. Some insects and at least one mammal, reindeer, are tetrachromats. They can see into the ultraviolet range.

Scientists could spend generations dissecting the brains of reindeer to discover the details of the brain states that are associated with reindeer vision; but, just having that knowledge isn't going to generate an instance of phenomenal ultraviolet in the experience of scientists who are familiar with the knowledge discovered by neuroscientists.

*What if we had talking reindeer? Would that help?*

It's not hard to imagine that scientists might someday genetically enhance the reindeer genome so that they might learn to speak a human language. Suppose that happened and that a committee of philosophers was formed to investigate whether reindeer know something about phenomenal color that humans don't know. We can imagine the following dialogue taking place between philosopher

Phil and Rudolph, the reindeer shop steward.

Phil: So, Rudolph, what does ultraviolet look like to you?

Rudolph: It looks something ultraviolet.

Phil: No, I mean what is it like to see something that reflects ultraviolet light?

Rudolph: That's it. That's what seeing something ultraviolet is like.

[Pause for a short lesson in how to talk about qualia.]

Phil: Let's take it from the top. Rudolph, what does ultraviolet look like to you?

Rudolph: It's like seeing phenomenal ultraviolet.

Phil: That's what it is. But what is it *like*?

Rudolph: It's just like seeing the quale of phenomenal ultraviolet.

Phil: I don't know what that's like! Please explain in terms I know about.

Rudolph: Well, you know what experiential color phenomena are like. You experience other colors. The experiential phenomenon we both call phenomenal ultraviolet is just another experiential color phenomenon; but, it's different from all the others.

Phil (thinking he may be making progress): Okay, so what's it like?

Rudolph: If you have to ask, you don't know.

Phil: I know that I don't know.

Rudolph: Well, as the most famous reindeer philosopher once said, knowing that you don't know is the beginning of wisdom. Suppose that doctors and scientists of the future find a way to insert genes crucial to Reindeer vision into humans; perhaps, by tacking them onto an RNA virus that infects retinal cells, inscribes the genes into the DNA of the human cells and then activates those genes so that infected humans start growing the fourth kind of cone. You'll then be able to see phenomenal ultraviolet. If all that ever happens to you, do you think you will learn anything from seeing phenomenal ultraviolet for yourself?

Phil: I'm not sure.

Rudolph: It was not a rhetorical question; so, let's figure it out. How do you make the transition from not knowing to knowing what experiencing phenomenal ultraviolet is like without actually learning anything at all?

Phil: Now *that* was a rhetorical question! I would pose the question thus: How do I explain the transition from not knowing to knowing what experiencing phenomenal ultraviolet is like without learning anything that threatens physicalism?

*Consequences of Churchland's Concession*

Significantly, accepting the intuition that discursive knowledge will not generate knowledge by acquaintance has consequences. It would mean that first-person phenomenology is not just the only way to discover the phenomenal facts, it's also the only way to discover which experiential phenomenon is associated with a

given neural phenomenon or brain state.

Howard Robinson (2012) develops the anti-physicalist consequences of this conclusion. Other philosophers have developed versions of physicalism that could survive Churchland's concession; for example, Robert J. Howell (2009; 2013) to be discussed below. However, it is not clear that identity theory physicalism, the form of physicalism favored by Churchland, can survive acknowledging the inability of knowledge by description to induce an instance of knowledge by acquaintance.

*How will science rescue identity theory physicalism?*

In a postscript, Churchland pins the defense of physicalism on the possibility that scientists might someday show that the physical is identical to the phenomenal.

> To be sure, the materialist cannot simply insist that Mary's neuroscientific knowledge truly constitutes a knowledge of sensations and their qualities: that would beg the question against the property dualist. Whether such identities do or do not hold is for our unfolding science to show in the fullness of time. But equally sure, neither can Jackson simply insist that neuroscience must *fail* to encompass qualia ... That would be to *assume at the outset* that materialism is false rather than to *show* that it must be false. (Churchland, 1997, 176)

*How likely is it that scientists will prove an identity theory?*

Scientists typically investigate the causal interactions between items under investigation; but, showing a causal relation between something material and something phenomenal would establish a non-identity relation. Something can't be identical to its own cause. Consequently, it doesn't seem at all likely to me that scientists will eventually vindicate identity theorists.

On the seemingly reasonable assumptions that knowledge by acquaintance is knowledge of experiential phenomenal and that the relevant knowledge by description is knowledge of neural firing patterns, quantum microtubular computations, wavefunction collapse or some other physical phenomena occurring in the brain, proving an identity claim would involve proving that something which exists in an experiencer dependent way is identical to something that exists in an experiencer independent way.

*Are there any neuroscientists actually working on that?*

Seriously, that is not a rhetorical question.

*Are there any philosophers working on that?*

By “working on that” I mean working to show that something which exists in an experiencer dependent way is identical to something that exists in an experiencer independent way.

### §3.3.2.5 Promissory Deductivism

In 1998 or thereabouts, Jackson switched sides with respect to the KA. In the all too brief “Postscript on Qualia”, he states:

> I now think that the puzzle posed by the knowledge argument is to explain why we have such a strong intuition that Mary learns something about how things are that outruns

> what can [be] deduced from the physical account of how things are. (Jackson, 1998/2004, 419).

According to Jackson, this intuition comes from the nature of sensory experience itself.

> … sensory experience presents itself to us as if it were the acquisition of information about intrinsic nature. But, very obviously, it is not information about intrinsic *physical* nature, so the information Mary acquires presents itself to us as if it were information about something more than the physical. This is, I now think, the source of the strong but mistaken intuition that Mary learns something new about how things are on her release. (Jackson, 1998/2004, 419).

I readily agree that sensory experience presents itself in such a way that we easily come to believe that colors are intrinsic properties of the objects we are looking at; but, I deny that the falsity of this belief is at all obvious. It takes a certain familiarity with the science of color vision to realize that physical objects are not colored in the way that a naively realistic experiencer might assume. Consequently, in my view, the strong intuition that Mary learns something upon her release derives from an inference to the best explanation drawn after comparing the phenomenology of color experience and the relevant scientific knowledge.

In any case, the question on which the evaluation of the KA turns is whether the intuition is *mistaken*.

One strategy for arguing that the intuition is mistaken would be to adopt color realism, the assumption that physical objects do, in fact, have the color properties that sensory experience presents them as having. If color realism is false, it is hard to escape the conclusion that Mary learns *something* upon her release even if it is only the nature of the (illusionary) color phenomenon associated with seeing various objects. Which versions of physicalism, if any, would be threatened by the falsity of color realism is another matter; but, one that would have to be investigated.

Jackson opts for a different strategy, that of assuming that the redness of phenomenal or qualitative red “can be deduced in principle from enough about the physical nature of our world despite the manifest appearance to the contrary that the knowledge argument trades on” (Jackson, 1998/2004, 417-418).

*Has Jackson introduced a third way of knowing into the discussion?*

Jackson acknowledged referencing two kinds of knowledge (or ways of knowing) in the KA, knowledge by description and knowledge by acquaintance. Now, to avoid the conclusion to which the KA might lead us, Jackson introduces knowing by deduction.

I am not trying to revive the allegation of equivocation that Churchland first raised against the KA. I think Jackson was right to claim that the KA was about the total knowledge that Mary had available to her before her release compared to the new knowledge she seems to acquire upon her release; and, I have no objection to granting Mary all the knowledge she might deduce from whatever she knows before her release.

*Does the knowledge gained by deduction include or generate an instance of knowing by acquaintance?*

We still have the problem raised in the previous section in connection with Churchland's concession that no amount of discursive knowledge will constitute an instance of knowing by acquaintance. If deductive knowledge doesn't include or generate knowledge by acquaintance, it is hard to believe that Mary doesn't learn anything upon becoming acquainted with experiential color phenomena upon her release.

When scientists attempt to deduce a consequence from a theory, the result is knowledge of what that theory predicts; but, it is not knowledge of what happens when an experiment is carried out. To obtain that empirical knowledge, the scientist must actually conduct the experiment to discover whether the theoretical prediction was correct.

It would seem that, even if pre-release Mary could deduce some idea of what experiencing color phenomena will be like, she won't know whether her prediction is correct until she sees for herself. So she will still learn something upon her release.

If the claim is that deductive knowledge will include or generate an instance of knowing by acquaintance, the claim is not just wildly implausible, it is extreme enough to be testable. We need only assume, as a null hypothesis, that the deductive instantiation of experiential phenomena does not occur; and, invite Jackson or a like minded philosopher to present the deduction that will induce an instance of *phenomenal ultraviolet*, in a normal human trichromat.

*Is that too tough a test?* Maybe. But, until it is met, it seems that Schrodinger was right.

> The sensation of color cannot be accounted for by the physicist's objective picture of light-waves. Could the physiologist account for it, if he had fuller knowledge than he has of the processes in the retina and the nervous processes set up by them in the optical nerve bundles and in the brain? I do not think so. (Schrodinger, 1958)

* * *

Until the deduction of phenomenal ultraviolet is actually presented, all we have defending a priori physicalism from the KA is ... *promissory deductivism*.

Prospects for a priori physicalism, never bright to begin with, were diminished greatly by a recent development in physics. It's been proven that, even with a complete microphysical description of a certain material, the question of whether that material possesses a certain macroscopic property, a spectral gap, is mathematically undecidable. (University College, London, 2015).

If knowing all the microphysical facts about a given material doesn't allow one to deduce whether it has a macrophysical property, there may be little hope of deducing whether a complex arrangement of material will or will not have some associated phenomenality. The prospects for being able to deduce what it would be like to experience that phenomenality also diminish.

In any case, until Mary can generate an instance of phenomenal redness by deduction from her theories about the brain activity associated with trichromatic vision, she will learn something after her release.

### §3.3.2.6 A Muted Acquaintance with a Revised Representationalism

Over time, Jackson developed an more elaborate defense of physicalism. In Jackson (2003) he indicated that he supported a priori physicalism instead of the more popular view that the phenomenal fact are necessitated by the physical facts but are not a priori deducible from them.

At the heart of Jackson's new version of representationalism[8] is the claim that the phenomenal redness that Mary experiences is not an instantiated property.

> There is a redness about sensing red ... We naturally think of the redness as a property we are acquainted with when we sense red and as the property Mary finds out about on her release. … Intensionalism tells us that there is no such property. To suppose otherwise is to mistake an intensional property for an instantiated one. (Jackson, 2003, 430)

Jackson goes on to say that the redness is only a feature of how things are represented as being and that it is not an argument against physicalism that “people are sometimes in states that represent that things have a non-physical property. ... What physicalists must deny is that such properties are instantiated.” (Jackson, 2003, 431)

Jackson briefly mentions first-person phenomenology. He says that, if a representational state is richly detailed enough and plays the right functional role, “we get the phenomenology for free” (Jackson, 2003, 438). However, it is not clear how that is supposed to happen.

Similarly, Jackson downplays the role of acquaintance in knowing first-person phenomenology. Officially he says the redness that Mary experiences "is not a feature one is acquainted with, but instead is a matter of how things are being represented to be" (Jackson, 2003, 432). However, in explaining his opposition to the idea that perceptual representation is nonconceptual, he argues that perceptual representation is inherently conceptual because representing that something is X "essentially involves discrimination and categorization, and that is to place things under concepts" (Jackson, 2003, 435).

Consequently, Jackson acknowledges that he might experience a particular shade of red for the first time and not know the term by which it is known; and, he denies that, when he later learns the term, "red-17", for that shade of red, he is learning a new concept. "It will simply be acquiring a term for something I already grasp" (Jackson, 2003, 435).

The situation with Mary is actually the reverse of the one Jackson just imagined himself being in. Mary knows all about the brain states associated with color phenomena; and, about the color terms typically used to describe the experiential phenomena associated with each such brain state. Only when she finally experiences phenomenal color for herself does she suddenly grasp that to

---

8 The representationalism of Jackson (1977) is quite friendly to sense data.

which the term refers.

That sudden grasping is learning from experiencing – knowing by acquaintance, in the traditional terminology. She becomes acquainted with an experiential phenomenon, tomato red, she had not previously experienced/grasped. So, Jackson's position requires that there be knowing by experiencing – acquaintance – to explain why, in certain circumstances, there is no learning by acquiring terminology.

### §3.3.2.7 The Transparency Argument

Jackson acknowledges the intuitively appealing conclusion of the KA, "... that Mary learns a new way in which certain items, in particular certain experiences, are alike." He then suggests that "... the best way to attack this contention – the 'new similarity' contention, as I will sometimes call it – is via representationalism about sensory experience". (Jackson, 2007, 53)

Jackson initially distinguishes minimal representationalism, the claim "that experience is *essentially* representational", from strong representationalism, the claim "that experience is *exhaustively* representational" (Jackson, 2007, 57). However, he later talks of weak representationalism as denying strong representationalism but making a claim that goes beyond minimalistic representationalism; namely, the claim that "the manner in which she [Mary] represents is an additional factor in making her experience the kind of experience it is" (Jackson, 2007, 63).

The argument for strong representationalism from minimal representationalism depends on two assumptions: the diaphanousness (transparency) of experience; and, the thesis of univocality (to be explained below). Significant for the moment is Jackson's claim that only strong representationalism can blunt the KA.

As might be imagined, there are any number of possible objections to such an extreme claim. At least some would be highly contentious; and, it is not my intention to resolve any of them; but, unless all these objections are resolved in Jackson's favor, his argument fails to blunt the KA.

#### *Objection 1: Even Strong Representationalism is Not Enough*

One might argue, as does Torin Alter (2007), that not even strong representationalism blunts the KA. Jackson concedes that Alter has successfully shown that weak representationalism is not sufficient to undermine the KA; but, denies that Alter has shown that strong representationalism would also fail.

#### *Objection 2: Denying that Experience is Essentially Representational.*

Minimal representationalism is supposed to be the uncontroversial claim that experience is essentially representational; but, that claim can be contested.

Jackson points out one way of doing that: asserting a projective theory of experience. We can reasonably assume that, in the causal story leading up to seeing a tomato, important events occur in the brain. At some point in this process, I am presented with an instance of phenomenal redness experienced as being *elsewhere* than in the brain. It is experienced as being on the surface of

the tomato.

Physical phenomena occur within a nervous system; but, correlated experiential phenomena are generally experienced as being outside the nervous system. I am *here*; but, the sensory phenomena instantiated in my experience are taken as being *out there*. That is what I'm calling projection.

Suppose a representative sample of experimental subjects look at an inkblot; and, that some see a bat where others see cat. Neither the bat nor the cat are actually in the inkblot despite being experienced as being there. Rorschach inkblots are designed to be ambiguous stimuli, which stimuli are organized by the brain into an image meaningful to the experiencer. “It is assumed that whatever the patient sees is a projection of his/her personality.” (Thomason, 2013)

However, Jackson doesn't show that projective theories of perception are false; only that they are incompatible with an element of his position, either the principle of transparency or the claim that experience is essentially representative. Thus, his argument is incomplete until it is shown that projective theories of experience are false.

### *Objection 3: Denying that Experience is Representational*

One may even see projective aspects in the theoretical approach philosophers take to describing the problem to be solved. According to Kobes (2007), the Philosopher's Projective Error is

> … a manifestation of our tendency, when reflecting on the relation between perception and the world, to treat what are in fact certain properties of perceptions as themselves perceived, or as properties of things perceived, or as represented in perception as properties of things perceived. (585)

The representational approach holds that experience has representational or propositional content. In the case of Mary and the tomato, the propositional content of the experience is the proposition *that the tomato is red*.

It is possible to deny that such a proposition is in the experience except as a projection of the philosopher's theory of experiencing. As I will later elaborate in presenting the Additional Abilities Hypothesis, one may take back the projection or *deproject* by taking responsibility for constructing the proposition in question.

In my view, I have the ability to notice ongoing experience and to comment on it. My commentary will usually include propositions such as “the tomato is red” or “the tomato looks red to me” or whatever; but, I construct these propositions in response to experiencing. I see no reason why one of these propositions (that the tomato *is* red) should be considered *the* one and only proposition in the experience; so, I deny that any of them form the propositional (representational) content of experience.

For Jackson's version of representationalism to be true[9], the proposition must be

---

9 Versions of representationalism in which the qualitative content of experience (either alone or in conjunction with the propositional or conceptual content of experience) makes an experience the experience it is may be more compatible with projectionist; but, if such a version of

in the experience *in an experiencer independent way*. In a projective theory of experiencing, the proposition is in the experience (or out there in the tomato) only if the experiencer puts it there.

### *Objection 4: Denying Color Realism*

Arguably, representationalism requires color realism. Lycan writes

> The representational theory seems to require color realism, on pain of circularity. In all the preceding discussion, color words such as “green” have been used to mean objective, public properties of physical objects. One could not (without circularity) explicate phenomenal greenness in terms of represented real-world public color and then turn around and construe the latter real physical greenness as a mere disposition to produce sensations of phenomenal greenness, or in any other way that presupposed phenomenal greenness. (Lycan, 2015)

Lycan goes on to suggest that representationalism does not require color realism because the representationalist might hold an error theory of color. It might seem somewhat peculiar for any representational theory to rely on a claim that experience misrepresents reality; but, be that as it may, Jackson's version of strong representationalism certainly seems to entail color realism by way of his *univocality thesis*.

> The question, What makes it right to use the word “square,” say, both to capture the nature of an object and to capture the nature of an experience? cries out for an answer. Representationalism explains this nonaccident by a certain kind of univocality thesis. To illustrate with the word “square”: it applies to something if and only if it has the property of being square; it applies to a visual experience if and only if the experience represents something as having the same property of being square. No special sense of “square” enters the story – to be designated “square*,” as it might be when philosophical perspicuity is important – in order to account for why “square” applies to visual experience. (Jackson, 2007, 54)

If the univocality thesis applies to color terms, the problem quickly becomes obvious. This version of representationalism assumes that there is only one sense of “red” and it applies both to the nature of objects and the nature of experiences. However, if the causal theory of perception is true, experiential phenomena such as redness occur *later* in a sequence of causal events that includes physical phenomena such as light reflecting off of surfaces of objects, light being refracted by transparent objects or volumes and light being emitted from objects.

Consequently, these physical phenomena can't be identical experiential color phenomena occurring later in the sequence of causal events. So, clearly, if any of the physical phenomena just mentioned are the reason why color realists say that objects have color properties, there are two senses of color terms in play.

At the very least least, Jackson's version of representationalism needs to incorporate a defense of color realism if objects are said to have color properties or a defense of an error theory if objects lack such properties.

---

representationalism were true, Jackson's argument would fail to blunt the KA. The strong representationalism required to blunt the KA assumes that experience has no content other than the propositional content.

### *Objection 5: Denying the Transparency (Diaphanousness) Thesis*

> Diaphanousness says that the properties of the object of experience determine without remainder the nature of the experience. It follows that if the object of experience is an intentional object, the experience's properties are one and all the properties of how things are being represented to be. Here I mean the experience's properties qua kind of experience it is. (Jackson, 2007, 60-61)

One may simply deny (as I do) that experience is a property bearer; and, that all talk about the properties of experience involves a category error.

At the cost of begging the question against those few philosophers such as Dennett who are eliminative as to first-person phenomenology, we can assume that there is first-person phenomenology; and, that our objectives are to describe and explain it. Now, it seems intuitively obvious that the items to be described are phenomena; and, that the properties of various physical objects are relevant to the task of explaining the occurrence of experiential phenomena; but, it's not so clear that there are any intentional objects that might be invoked to explain experiential phenomena. The experiential phenomenon of tomato red, say, is the intentional object whose occurrence is to be explained.

Assuming arguendo that the transparency thesis could be restated so that physical phenomena and/or the properties of physical objects are the only determinants of the nature of an experience, the “without remainder” clause of the restated thesis would be still be vulnerable.

Suppose I am looking at a bistable image; for example, the gestalt drawing of the vase/faces. Whether I am currently seeing the faces or the vase, I can always choose to see the other image. The properties of the object I am looking at doesn't change. It's still a drawing in black ink on white paper. None of its pixels move. None of its properties change.

But the image flips from one view to the other in response to my intent; or, so it seems to me. If my intent helps to determine what I see, the transparency thesis is false.

Ambiguous figures such as the inkblot can also challenge the transparency thesis. If the brain is spontaneously organizing the ambiguous stimulus into a definite image, the properties and activity of the brain contribute to what I see.

### *Objection 6: The Unit of Analysis Is the Phenomenon*

> However, the key question for whether representationalism undermines the knowledge argument is … whether the new kind of experience Mary has when she first sees red is a reason for her to enlarge the range of properties she holds to be instantiated in our world. (Jackson, 2007, 63)

Nowhere does Jackson give an impression of taking an eliminative stance toward first-person phenomenology. He speaks about the incredible richness of first-person phenomenology; about getting it for free from the right kind of representational state; and, about the transparency thesis being a thesis about the phenomenology of perceptual experience.

But, given that there is first-person phenomenology, one would naturally expect that, in any attempt to describe or explain first-person phenomenology, the unit

of analysis would be the phenomenon. Furthermore, first-person phenomena must exist while they are occurring; otherwise, there would be no first-person phenomenology at all.

The new kind of experience Mary has when she first sees red is the kind of experience it is because of the kind of first-person phenomenon she experiences; and, it's new because she experiences that first-person phenomenon for the first time; so, it's *new to her* not new to the world. Consequently, Mary has good reasons for enlarging the range of experiential (first-person) phenomena *instantiated in her experience* even if she doesn't have any reason for enlarging the range of physical (third-person) phenomena she holds to be *instantiated in the world*.

Since Mary already knew everything she could learn from her lessons about any and all physical phenomena related to vision, she learns something new. She knows something about the way experiencing is. She now has information about first-person color phenomena. If that counts as phenomenal information which is not physical information (and I don't see why it wouldn't), physicalism is refuted. “The knowledge argument works. There is no way to grant the hypothesis of phenomenal information and still uphold materialism.” (Lewis, 1988, 90)

In view of the foregoing considerations, it seems clear that the KA loses none of its force when phenomenal redness is considered a phenomenon rather than a property. On the other hand, strong representationalism can no longer blunt the KA simply by saying that an experiential phenomenon is not an instantiated property.

Jackson would need to find some other way to blunt the KA. Could a defense parallel to the strong representationalism that Jackson offers be constructed around the idea that phenomenal redness is an experiential phenomenon? He would not be able to claim that experiential phenomena are not instantiated phenomena as that would undermine the non-eliminative stance he appears to take.

* * *

Strong representationalism is a strong claim with several points of vulnerability. I doubt that the last word has yet been written concerning attacks on these vulnerabilities or defenses that strong representationalists may have. At the moment my point is that all of these attacks must be defeated or Jackson's attempt to blunt the KA fails.

### §3.3.2.8 Updating Type-Type Identity Theory

Jackson (2012) next presents an updated version of a type-type identity theory; arguing that the type-type version of identity theory is “the version of choice for those who favor some version or other of functionalism”.[10] (Jackson, 2012, 150)

---

10 Jackson does not explicitly present the identity theory he discusses in this paper as his own position; but, as we've seen previously, his explanation for phenomenality seems to be that it happens “for free” when a rich enough representational state plays the right functional role. Assuming that a representational state is a representational state of the brain, his position

Although this essay doesn't mention the KA, one must wonder whether this version of identity theory can withstand attacks based on the KA.[11] Identity theory physicalism is a strong claim.

> Physicalism about consciousness is a strong claim. It isn't the relatively anodyne claim that conscious states are closely connected with what is going on in the brain, in the way that smoke is connected with fire. Rather physicalism says that conscious states are brain processes, in the way that water is $H_2O$. Smoke is caused by fire. But water isn't caused by $H_2O$ – it is $H_2O$. Similarly, according to physicalism, pain isn't caused by a brain process – it is a brain process. (Papineau, 2008, 57)

As the strongest possible version of physicalism, identity theory physicalism would be the most vulnerable, if any form of physicalism is vulnerable, to challenges based on the KA.

### §3.3.2.9 Observations and a Plan for Moving Forward

Some clarification of the criteria employed to judge the success or failure of the KA is sorely needed.

When Jackson recanted, he may have altered the criterion of success from whether something escapes the physicalist's story to whether Mary learns something new about *the way things are*. In my view, the question has always been whether she learns something about the way experiencing is rather than something about the way things are.

In any case, it has never been clear what counts as “escaping” the story told by physicalists. With so many versions of physicalism in circulation, it seems reasonable to suppose that more would be required to escape from some accounts than from others. Although I count myself as a proponent of the KA, I feel no obligation to escape from all possible versions of physicalism before claiming to have successfully escaped the story told by those physicalists who are (by my count) phenomenon monists.

In my view, Churchland glimpsed but did not cleanly identify the question on which the debate turns. The question Churchland actually posed – *whether sensory qualia form a metaphysically distinct class of phenomena beyond the scope of physical science* – is actually two questions conflated. Interestingly enough, the first is the same question that Searle asked (in another context).

1. *How many fundamentally distinct kinds of phenomena are there?*

If we answer “Two” based, let us say, on a claim that experiential phenomena constitute a class of phenomena ontologically distinct from physical (physical$_1$) phenomena, we face the next question.

2. *Is explaining the origin of experiential phenomena not identical to*

---

becomes a simple variation on the argument scheme he discusses concerning causal roles.

11 Interestingly, another essay (Hill, 2012) in the same volume in which Jackson's essay appears compares type-type (or central state) identity theory and representationalist theories with respect to their ability to account for pain, “the paradigmatic qualitative state”. (p. 130, fn 5) Hill's conclusion is that, while the debate is far from over, representationalist theories have the advantage at the present time.

*physical$_1$ phenomena beyond the scope of physical science?*

It is possible to answer “Yes” to the first question independently of one's answer to the second. We will have to wait and see whether – and how – scientists are able to explain experiential phenomena; but, while awaiting further bulletins from the frontiers of science, I propose to adapt the KA so that the criterion of success/failure depends on answering the first of the two questions that Churchland conflates. I will call the question around which the debate swirls the Churchland/Searle Debate Focusing Questions:

> **[DFQ-1] How many fundamentally distinct kinds of phenomenon are there?**

Further questions are only reached if one's answer to [DFQ-1] is “Two”. Such questions would include

> **[DFQ-2] Is explaining the origin of experiential phenomena not identical to physical phenomena beyond the scope of physical science?**

However, answering [DFQ-2] would require unpacking it; and, deciding whether having two fundamentally distinct kinds of phenomenon is consistent with physicalism, or dualism or both.

### §3.3.2.9.1 Setting [DFQ-2] to One Side

I will shortly propose a slightly weakened version of KA, that targets those who would answer [DFQ-1] with “One”. These would include identity theory physicalists and eliminative materialists. Before doing so, however, I want to say a few words about why I will set aside the second of the two questions that Churchland conflates, the question I'm labeling [DFQ-2].

It's not clear what is within the reach of physical science; but, one reasonable interpretation is that what physical scientists have explained is within the reach of physical science. One may then extend this thought, saying that scientists may be able to explain experiential phenomena some day.

However, scientists typically explain things by discovering the causal relations between the items under investigation. This creates a problem. Churchland's argument has always been that advocates of the KA haven't ruled out the possibility of an *identity* claim between experiential and physical phenomena; but, if experiential phenomena are the effects of physical causes they can't be identical to those causes.

In any event, when considering the possibility of eventually getting a scientific explanation of experiential phenomena, it is assumed that the explanans of a successful physical explanation will invoke only items that are physical$_1$ – physical in the strict, mind independent sense. However, if experiential phenomena are mind dependent – depending for their existence on the experiencing subject – they can't be physical$_1$ phenomena.

Now, as we saw in considering Searle's attempts to avoid being considered a dualist, one might simply expand the definition of “physical” so that physical$_2$

includes mind dependent, experiential phenomena *caused by* something (e.g. the brain) that is itself physical$_1$. However reasonable such an expansive definition may seem when contemplating the ever increasing explanatory power of science, a successful scientific explanation of experiential phenomena would not show that they are physical$_1$ phenomena, only that they are physical$_2$ phenomena.

There would still be two kinds of phenomenon.

There would likely be considerable controversy as to whether having two fundamentally distinct kinds of phenomena, physical and experiential, is consistent with physicalism. There are any number of physicalistic perspectives and the details of each perspective would have to be specified before one could decide whether it was compatible with recognizing two fundamentally distinct kinds of phenomena.

My concern is that classifying a philosophical system as monistic or dualistic should be simply a matter of counting.

In my view, if experiential phenomena are not identical to physical$_1$ phenomena, there are two fundamentally distinct kinds of phenomena. I would consider that to be phenomenon dualism; *irregardless* of whether some philosophers find reasons satisfactory to themselves for calling the resulting perspective a kind of physicalism. I simply find it impossible to believe that having two fundamentally distinct kinds of phenomena should constitute phenomenon monism.

In my view, a given philosophy may constitute both physicalism and phenomenon dualism; but, such a claim would likely provoke controversy. That's one reason I will set that and other issues aside for the moment. I will focus on whether the next generation knowledge argument, KA:TNG, provides us with any reasons for answering [DFQ-1] with “Two”.

### §3.3.2.9.2 Embracing the Parody

To review: Churchland (1989) rejected Jackson's attempt (1986) to avoid the charge of equivocation that Churchland (1985) leveled against Jackson's original (1982) formulation.

I happen to think that Jackson succeeded in avoiding equivocation simply by following Churchland's lead and distinguishing the two modes of knowing at issue. The essential question is whether Mary learns anything when she becomes acquainted with *tomato red*, the experiential phenomenon.

Among his other objections, Churchland offered a parity of reasons objection to KA. The objection has the form of a reductio; but, his tone in presenting it suggests a attempt at parody, heaping further scorn on an argument that he rejected for other reasons. Nevertheless, there is something insightful about his parity of reasons objection. So, I want to try to rescue the insight from the parody so that we might see whether it supports or undermines the KA treated as valid.

Churchland's objection is that the KA is indiscriminately anti-reductionist. Since Churchland is an identity theorist and a reductionist, he naturally thinks of that as a bad thing. I am a non-identity theorist and an anti-reductionist; so, I think

having an indiscriminately anti-reductionist argument is a good thing. It's like having a broad spectrum antibiotic. It's good for whatever ails you.

Specifically, Churchland argues that if the KA is effective against reductive materialism it would be equally effective against substance dualism; but, how many substance dualists offer a reductive account of experiential phenomena? Descartes doesn't. In his view, sensations – anything would fall under the narrow view of qualia due to Lewis – were the result of the interaction of body and soul. Thus, no experiential phenomena would be identical to either the activity in the body or the activity in the soul at the time of the experience.

It is hard to imagine how Descartes' non-reductionist position would be adversely affected by an argument that is indiscriminately anti-reductionist.

Nevertheless, Churchland's point is that in parallel cases we would expect the same anti-reductionist outcome. The same argument may be applied against the substance dualist who makes an identity claim analogous to the one made by the identity theory physicalist.

I don't know of any actual substance dualists who claim that the soul/experience relation is that of identity; but, for the purposes of a thought experiment, that doesn't matter. The identity theory physicalist is committed to the claim that the color of an afterimage is identical to the activity going on in the brain at the time of the experience. So, we simply imagine someone claiming that the color of an afterimage is identical to the activity going on in the soul at the time of the experience.

Adapting Churchland's point to the case of a formally valid argument, we should be able to substitute lessons in spirituality for lessons in physicality. We simply assume that, during her confinement, Mary learns all the spiritual facts about human color vision that her lessons can give her. The rest of the story is the same. Upon her release, Mary encounters a ripe tomato and exclaims, “Ah! So, that is tomato red” or something like that.

Now, once we bring the item that Mary encounters upon her release, tomato red, under the concept of *phenomenon* rather than the concept of *property*, we know that the phenomenon is an appearance. We then conclude that an appearance is not identical to the underlying reality of which it is merely an appearance – regardless of whether a materialistic or spiritualistic ontology is alleged to explain that appearance.

In the case of identity theory materialism, attempting to identify the experiential phenomenon of, say, tomato red with the corresponding physical phenomenon, NFP-4738, is doomed to failure. An experiential phenomenon occurs in an experiencer dependent way; so, it can't be identical to a physical phenomenon that occurs in an experiencer independent way.

Suppose there were an identity theory spiritualism which held that tomato red is identical to the spiritual phenomenon occurring in the soul at the same time as the experience. In an exactly parallel case, where the spiritualist held that spiritual phenomena occur in an experiencer independent way, we could make the parallel counterargument: An experiential phenomenon occurs in an experiencer dependent way; so, it can't be identical to a spiritual phenomenon

that occurs in an experiencer independent way.

So it seems that the KA is an indiscriminately anti-identity argument. To the extent that a philosophical position is reductionist because of an identity claim, that position will be in conflict with the KA.

Thus, one more reason why the next generation knowledge argument will explicitly target identity claims is simply that the original argument did so implicitly.

### *§3.3.3 KA, The Next Generation*

In this section, I propose to reformulate the KA so that success/failure depends on answering [DFQ-1]. Specifically,

1. Success for opponents of the KA consists in showing that "one" is the correct answer; and,
2. Success for advocates of the KA consists in showing that "two" is the correct answer.

The situation is this.

> **[KA:TNG-1] Mary (before her release) has complete knowledge by description of everything physical in any way related to color vision; knowledge by deduction whatever is deducible from whatever she knows by description; and, may be credited with knowledge acquired by other ways of knowing – except knowing by acquaintance.**
>
> **[KA:TNG-2] Mary (after her release) becomes acquainted with experiential phenomena related to color vision.**

Nothing much follows from these premises alone; so, philosophers of whatever persuasion will have to work for their conclusions. There may be any number of possible conclusions toward which philosophers might aim depending on their objectives; but, my objective is to show that experiential phenomena *survive* the story told by eliminative materialists and *escape* the story told by identity theory physicalists. Consequently, I propose a trio of alternate conclusions, one each for eliminativists, identity theorists and non-identity theorists to aim for.

> [Where P = "physical$_1$ phenomenon" and Q = "experiential phenomenon"]
>
> **[KA:TNG-C1] There are no experiential phenomena at all.**
>
> **$\neg(\exists x)(Qx)$**
>
> **[KA:TNG-C2] For any experiential phenomenon, there is a physical$_1$ phenomenon to which it is identical.**
>
> **$(\forall x)[Qx \rightarrow (\exists y)(Py \ \& \ x = y)$**
>
> **[KA:TNG-C3.1] There is at least one experiential phenomenon that is not identical to any physical$_1$ phenomenon.**
>
> **$(\exists x)(Qx \rightarrow (\forall y)(Py \rightarrow \neg(x = y)\ )\ )$**

Alternately, non-identity theorists may aim for a stronger conclusion:

> **[KA:TNG-C3.2] An experiential phenomenon is not identical to any physical$_1$ phenomenon.**
>
> **$(\forall x)(Qx \rightarrow \neg Px)$**

The story of Mary stays pretty much the same with one additional specification. As before, we assume that Mary is a brilliant student able to absorb all the knowledge that her lessons can give her; that she has become a brilliant scientist able to conduct her own research via monochromatic interfaces to the external world; and, that she is an ideal reasoner able to deduce whatever is deducible from whatever she knows by description. In addition, I will assume that Mary had a wise old grandmother who passed on various pearls of practical wisdom accumulated over the course of a lifetime. Let us imagine Mary's grandmother on her deathbed summoning up her last reserves of energy to email Mary saying, "Just because all the other philosophers in the world assume their way across an epistemological gap doesn't mean you have to do so as well".

* * *

The epistemic situation in which the debate takes place is very simple. It only takes one white crow to prove that not all crows are black; and, Mary gets to pick her crows. Due to her lessons, Mary knows that she can get two color experiences for the price of one tomato; so, she picks her two crows accordingly. As candidates for being experiential phenomena not identical to any physical phenomenon, she picks the phenomenal colors, *tomato red*, that she experiences when looking at a ripe tomato, and *otamot green*, the color of the greenish afterimage she sees after staring at her tomato in total fascination for too long and then looking at a brightly lit white wall.

In sections §3.3.3.1 and §3.3.3.2 I'll comment on the premises of KA:TNG. In §3.3.3.3 I will revisit the Nagel/Churchland argument and show that KA:TNG is immune to this objection. Following that, in §3.3.3.4 and §3.3.3.5, I'll review existing identity theories for any light they may shed on Mary's plight.

Finally, in §3.3.3.6, I'll focus on evaluating the justifications for a claim of physical/phenomenal identity.

### §3.3.3.1 What Does Mary Already Know?

My first premise clarifies Jackson's original first premise in two respects.

First, I've incorporated a point Jackson made explicit only after changing his stance toward the KA: that we must allow Mary to know by deduction as well as by description. In my view, the KA has always been about whether Mary learns anything by becoming acquainted with experiential color phenomena; so, this clarification seems reasonable to me.

Secondly, my first premise avoids an (alleged) ambiguity in Jackson's first premise concerning what counts as a physical fact. In Jackson's original formulation of the first premise, Mary has all the physical facts. This has been challenged by those who allege that there are subjective physical facts.

Robert Van Gulik describes the problem well.

> If one is thinking of physical knowledge as solely third-person knowledge of the sort one might get from studying physical theory, then physicalism's claim to completeness might seem to preclude there being any subjective facts knowable only from a given experiential perspective. … Thus, if 'physical facts' means 'physical theory facts', then subjective facts – if any exist – would seem to fall outside that range.
>
> However, if one reads 'physical facts' in a broader way to mean all the facts 'that obtain *in virtue of* physical processes' or 'that are *realized by* underlying physical structures', then it is far less clear that no such facts might be subjective in the sense of being perspectivally restricted in their knowability. (Van Gulik, 2004,388)

Clearly, if one expands the definition of “physical” to include what Mary did not know during her confinement, one could claim that there is a contradiction in the first premise in the KA. Mary is said to have all the physical facts but she doesn't have all the facts that might be considered physical under an expansive definition of “physical”.

Still, it is obvious that the facts that are allegedly physical only under an expanded definition of “physical” are facts that could not be taught in lessons.

In my view, it is very doubtful that Jackson incorporated such a contradiction into the KA; and, even if there was an ambiguity to begin with, it was removed when Jackson acknowledged Churchland's point that the first premise involved knowledge by description and the second premise involved knowledge by acquaintance.

In any case KA:TNG avoids a similar challenge because it explicitly says that Mary has complete *knowledge by description*. Does she have a complete set of physical facts? Well, not if “physical” is being used in an expansive sense; but, arguably, having a complete knowledge by description of everything physical constitutes having all the physical facts in the stricter sense of “physical” – all the $\text{physical}_1$ facts.

Now, the subjective physicalist may reply that, by weakening the premises of KA:TNG to avoid the appearance of a contradiction in the first premise of KA:TOA, I have made it too weak to refute physicalism because it does not refute subjective physicalism.

My response is that KA:TNG only attempts to refute forms of physicalism that allege that there is only one kind of phenomena, $\text{physical}_1$ phenomena.

If subjective physicalists have incorporated into their notion of the physical a second kind of phenomenon, one knowable only by acquaintance, they recognize two kinds of phenomena. It doesn't matter whether one has some reason for claiming that facts about experiential phenomena are $\text{physical}_2$ facts given some expanded definition of “physical”. If experiential phenomena are not $\text{physical}_1$ phenomena, then there are two kinds of phenomena under discussion.

#### §3.3.3.1.1 Subjective Physicalism

Robert J Howell (2013, 2009) holds that there are two kinds of physical facts, objective and subjective. According to Howell, Mary before her release only

knows the objective physical facts concerning color vision because that's all that her lessons can teach her.

She does not know the subjective physical facts concerning color vision until she leaves her room because one can not grasp the subjective aspect of a brain state until one's brain instantiates that brain state. Thus, according to the subjective physicalist, Mary doesn't know all the physical facts about physical objects because she doesn't know the experiential phenomenon associated with a given brain state about which she already has complete *objective* knowledge.

Howell agrees that Mary learns something when she first encounters a red rose; but, argues that it is a fact about the world, about the brain state instantiated by looking at the rose. This doesn't show that Mary learns about a new property unknown to physical science; only that humans are "creatures that have states that can only be fully grasped by occupying those states". (Howell, 2015, 37)

Howell doesn't exactly say that Mary come to know an old property in a new way. He says that some physical properties have *aspects* that are unknown to physical scientists but which Mary grasps by instantiating the relevant brain state. "I have appealed to the existence of “aspects” of properties which are not themselves properties." (Howell, 2015, 37).

It seems likely that the physical state that can only be fully understood by being in that state could be described in terms of physical phenomena such as neural firing patterns. It also seems likely that an aspect of a physical property that can only be grasped from the first person perspective would be an experiential phenomenon.[12]

Just as I would say that an experiential phenomenon is not identical to the physical phenomena with which it is correlated, Howell would have to say something analogous. An aspect of a property which is not itself a property could not be identical to a physical property of which it is but an aspect.

Consequently, Howell seems to have a non-identity theory of some sort; and, assuming (as seems reasonable) that we can translate Howell's theory into the language of phenomena, it seems clear that Howell recognizes two distinct kinds of phenomena: physical phenomena such as neural firing patterns and what I call experiential phenomena but which he would call the subjective aspects of some physical properties.

It is instructive to compare how Howell's subjective physicalism fares against the original KA with how it fares against KA:TNG.

A subjective physicalist could argue that the first premise of the original KA is invalid because it assumes that Mary has all the physical facts when it is clear that she doesn't have any of the subjective physical facts about color vision. No such objection is possible against KA:TNG. It claims that before her release Mary has only the facts that her lessons can teach her (together with what she can deduce there those facts); and, Howell admits that Mary can not acquire what I call phenomenal facts or what he calls subjective physical facts without

---

12 It is not entirely clear how we are to understand *grasping* an aspect; but, it seems natural to me to regard grasping as acquaintance traveling incognito.

instantiating the relevant brain state.

How does subjective physicalism fare against the conclusion to which the original KA would lead us, the conclusion that the phenomenal facts escape the physicalists story? The subjective physicalist can simply reply that the phenomenal facts do not escape the story being told by subjective physicalists.

Using the standard of success Jackson adopted after recanting yields a more ambiguous result. Does Mary learn something new about the way things are?

Jackson (after his recant) would say that Mary (after her release) does not learn anything new about the way things are; and, I'm inclined to agree. She does not learn anything she couldn't learn in her lessons from physical scientists about the arrangement of physical objects such as neurons, the properties of those objects or the (objective) physical phenomena (e.g. neural firing patterns) associated with those objects.

The subjective physicalist would reply that Mary learns which phenomenal facts are associated with a given neural firing pattern; and, that such a fact is a subjective physical fact about a given arrangement of neurons and their activity.

I have some sympathy for the claim of the subjective physicalist; but, it seems to me that Mary already knew about all of that. Before her release, her Concordance of Terminology and Phenomenology would also tell her which words were typically used by humans to describe the experiential phenomena associated with a given arrangement of neurons and their activity. After her release Mary knows which experiential phenomena is associated with a given neural firing pattern and is referred to by those words listed in her catalog.

I would say that Mary learns something new about the way experiencing is. The subjective physicalist might say that Mary learns something new about the way things are or, at least, about what they can do. It's easy to imagine someone else coming along and saying that Mary learns something about the way things are *and* about the way experiencing is.

Consequently, it is not clear whether subjective physicalism would meet the standard of success advocated by Jackson after his recant.

None of this ambiguity matters with respect to KA:TNG because it targets the claim that there is only one kind phenomena; and, assuming that my translation of Howell's terminology into my own is sound, subjective physicalists hold that there are two kinds of phenomena. Consequently, KA:TNG is not an argument against subjective physicalism; so, the fact that subjective physicalism “survives” KA:TNG is not an objection to either KA:TNG or subjective physicalism.

This possibility illustrates the final ambiguity involved in the original KA that KA:TNG avoids, the question of whether one must establish that having two distinct kinds of phenomenon constitutes dualism before one escapes from any form of physicalism that holds that there is only one kind of phenomenon.

Non-identity theorists succeed by showing that KA:TNG establishes that there are two kinds of phenomena. It is not required that they also establish that having two kinds of phenomena constitutes phenomenon dualism (or some other form of dualism) instead of or in addition to constituting some form of

physicalism that is neither eliminative materialism nor identity theory physicalism.

Of course, I would think it absurd to hold that subjective physicalism is not a form of phenomenon dualism. I am reminded of a running joke from *Newhart*, an American TV sitcom in which comedian Bob Newhart plays an innkeeper in Vermont. Among the local residents are three brothers who make frequent appearances at the inn. Each time they appear, one brother, Larry, introduces the group by saying, “Hi, I'm Larry; this is my brother Darryl, and this is my other brother Darryl”. A good joke but a bad philosophy. Obviously, Larry has two brothers even though, for some obscure reason, they each have the same first name. Similarly, in my view, having two distinct kinds of phenomena would still constitute phenomenon dualism even if someone were to introduce their position by saying “I have two kinds of phenomena, these (objective) physical phenomena and those (subjective) physical phenomena”.

Howell, would undoubtedly disagree that subjective physicalism incorporates phenomenon dualism, constituting a dualistic form of physicalism. He takes great pains to distinguish his position from property dualism (Howell, 2009). I don't think much of those defenses; but, that is beside the present point: subjective physicalists answer [DFQ-1] by saying “Two”.

Consequently, assuming I'm right that subjective physicalism holds that there are two kinds of phenomena, KA:TNG does not attempt to refute subjective physicalism.

#### §3.3.3.1.2 Perry and the Assumption of Identity

John Perry is a antecedent physicalist attempting to defend his version of physicalism from three neo-dualistic arguments: the KA, the Zombie argument and the modal argument due to Kripke; but, I will only be concerned about his response to the KA.

Perry begins with an acknowledgment of the situation in which we find ourselves while considering our experiences. One wonders how *this* (some experiential phenomenon) could be identical to *that* (some physical phenomenon occurring in the brain at the same time). Intuitively, the very idea seems absurd.

> This feeling is what I will call the “Ewing intuition,” and the argument based on it, the “experience gap argument”: *this* could not be *a brain state*, because the gap between what it is like and what brain states are like is simply too large. (Perry, 2001, 4)

Perry then asks the question around which his argument revolves:

> ... can we really make sense of the thought that this feeling, this aspect of what goes on inside me that makes it a toothache or a headache or the smell of a gardenia or the taste of turnips, is an aspect of my brain that someone else could, in principle, see? (Perry, 2001, 9)

He then states "I will argue that we can" but doesn't actually argue *for* the identity claim. He argues *against* the anti-identity implications of the KA (and the other neo-dualistic arguments).

Perry describes his dialectical situation as that of someone already committed to

physicalism; hence, his term for his position, *antecedent physicalism*. The question the antecedent physicalist faces is whether identity theory physicalism remains possible in the face of the anti-identity implications of the neo-dualistic arguments against which Perry is defending his position.

He acknowledges the stakes clearly,

> If physicalism cannot accommodate the subjective character of experience, one must either give up physicalism or deny the subjective character of experience. (Perry, 2001, 27)

However, the antecedent physicalist rejects eliminativism and simply attempts to provide an account of

> ... subjective characters on the assumption that they are physical. Then, and only then, do we look at the neo-dualist arguments to see if they point out some inadequacy or hidden contradiction in our account. (Perry, 2001, 27)

I have no objection to a philosophers who specifies the details of his or her own position before, after or while engaging opposing arguments. I am, however, extremely skeptical of any philosopher who fails to state a case for his or her own position.

Ironically, Perry himself alludes to the tactics of defense attorneys; but, a legal analogy only illustrates the weakness of Perry decision not to present an affirmative case on behalf of the identity claims he is defending.

Suppose that Jack is accused of being the driver in a hit-and-run vehicular homicide in Wilmington, North Carolina. The prosecution presents its case by presenting witnesses who provide testimony. Let's say that a witness, Sally, claims to have seen the driver and recognized him as being Jack. After direct examination of the witness by the prosecutor, the defense attorney, Jill, attacks the credibility of the witness. Jill gets Sally to admit that the accident occurred at night about 50 feet from where Sally was standing and that she only caught a glimpse of the driver while he was momentarily illuminated by a streetlight while fleeing the scene at high speed.

Attacking the credibility of the witness is a legitimate tactic of trial attorneys. Prosecutors and defense attorneys may each do this to witnesses for the other side.

In the case at hand, Jill's aim is to undermine the jury's willingness to believe that the prosecutor has proven a crucial element of the state's case, that Jack was driving the car at the time and place of the homicide. Perry's defense of antecedent physicalism resembles this part of a defense attorney's actions on behalf of his or her client.

Eventually the state rests its case. Technically, the defense could rest its case immediately thereafter without putting on any witnesses of its own; but, a defendant would have to be very foolish to require his attorney to do so.

Suppose that Jack has a witness, Sam, who places Jack in Barstow, California, at the time of the homicide. He would naturally want Jill to call Sam to the witness stand and elicit Sam's testimony that Jack was 2500+ miles away from the scene

of the crime at the time of the crime.

On the basis of her attacks on the credibility of the prosecutor's witnesses, Jill might argue to the jury for a negative proposition, "The prosecution has not proven its case beyond a reasonable doubt".

On the basis of Sam's testimony Jill would argue for an affirmative proposition, "My client was elsewhere and could not have committed the crime."

Perry does not present an affirmative case for the identity claim. The closest he comes is to argue that alternatives to identity theory physicalism (supervenience physicalism and realization physicalism) each have problems that identity theory physicalism avoids.

In my view, this isn't good enough. I can easily agree that identity theory physicalism is the least problematic form of physicalism extent while still arguing that the problems it has are insurmountable.

So, the failure to present an affirmative case for the identity claim(s) being made is a serious weakness of Perry's argument.

It gets worse. Perry's defense depends on the assumed identity being true.

> If subjective characters are physical aspects of experiences, as the antecedent physicalist maintains, then if Mary knows all of the physical facts, she will know about subjective characters. (Perry, 2001, 98)

Logically, when the antecedent of a conditional claim is false, the conditional itself is considered true; albeit, vacuously so. If no argument for the truth of the identity claim is offered, one might reasonably consider it potentially false and the conditional claim itself potentially vacuous.

Resting the defense of identity theory physicalism on a potentially vacuous conditional strikes me as a desperate strategy; but, I wouldn't expect an identity theorist to give much credence to a non-identity theorist's view of the dialectical situation in which we find ourselves.

With these preliminary remarks about Perry's way of presenting his case, I'll turn my attention to the specific defenses he offers against the non-identity implications of the KA.

#### §3.3.3.1.2.1 The Source of the New Knowledge

It is widely agreed that, when she is released from her confinement and first encounters a ripe tomato, Mary will say something like, "*That* is tomato red".

The puzzle of the KA is that we can easily imagining Mary learning and using color terminology while still confined to the Jackson room. For example, she might say either of these two statements:

> [M-1u] Tomato red is the experiential color phenomenon associated with looking at a tomato
>
> [M-1m] "Tomato red" is the name people with normal color vision give to the experiential color phenomenon associated with looking at a tomato

These two statements differ in that the first is a use example and the second is a mention example.

There may be some dispute as to whether Mary is entitled to state [M-1u] or any use example. Employing a word in the *use* case involves referring to whatever the term refers to; but, before her release, Mary doesn't know what the term refers to.

Mary knows that the term is used to refer to a particular experiential phenomenon she has never experienced although others have; but, is merely propositional knowledge about the term. She doesn't know -- is not acquainted with -- the experiential phenomenon itself.

The identity theorist might say that a term may be a referring term even though one does not know what precisely it refers to. We can construct terms like "NFP-4738, the neural firing pattern associated with experiential phenomenon, tomato red". We may not know what "NFP-4738" refers to; but, that can be remedied, scientists can go looking for it and will, it is hoped, eventually find out which neural firing pattern "NFP-4738" refers to.

Attempting to fix the reference of a term this way doesn't suffice to make it a referring term, however. The neural correlate of tomato red may not be a neural firing pattern as that term is currently understood. We might find that the neural correlate of tomato red is a quantum microtubular computation instead of a neural firing pattern.

In the present case, though, it is known that there is something to which the term "tomato red" refers; so, we don't have the problem that it might turn out to be a non-referring term.

Nevertheless, it's not clear that Mary is entitled to state [M-1u] instead of [M-1m] because it's not clear that she can make a reference. Phenomenal concept terms are sometimes said to refer directly and immediately; so, if terms for experiential phenomenon have this feature, perhaps one doesn't actually refer unless one knows – in the acquaintance sense – what one is referring to.

The matter of direct reference for terminology for experiential phenomena is certainly a controversial one; but, Perry seems committed to claiming that, before her release, Mary is entitled to state either [M-1u] or [M-1m]; so, he should provide some justification for imagining her stating the former rather than the latter.

My point is that, prior to her release, she only has the knowledge implied by [M-1m]. She knows how to use the terminology. She does not know (in the acquaintance) sense the referent of her terminology.

Now, I hasten to concede that Mary may become very proficient in the use of "tomato red" and other color terms while still in confinement; meaning, that she can learn to employ those terms correctly during conversation.

Mary could participate in any number of debates on Internet mailing lists concerned with the philosophy of consciousness. She might discuss the relation between first-person phenomenology and the correlated third-person phenomenology of neurological events. Let us concede that she is so proficient in

using color terminology that she is never 'outed' as someone who doesn't know what she is talking about.

She appears knowledgeable because she knows how to employ the terminology; but, she does not know what she is talking about because she does not know – is not acquainted with – the referents of her color terms.

Similarly, I use *phenomenal ultraviolet* as a name for the experiential phenomenon tetrachromatic organisms (e.g. reindeer) experience when they look at something that reflects “ultraviolet” light. I may employ the term during conversations, including technical discussions of psychophysical correlations and the brain/experience relation; but, I deny knowing what I am talking about when I talk about phenomenal ultraviolet.

I do not know what I am talking about because I am not acquainted with the experiential phenomenon to which the term would refer if spoken by a talking tetrachromat.

In any case, we can imagine Mary making two analogous statements when, after her release, she encounters a tomato.

> [M-2u] Tomato red is *that* experiential phenomenon
>
> [M-2m] "Tomato red" refers to *that* experiential phenomenon

While [M-2u] uses a term that [M-2m] only mentions, this time there is no doubt that Mary is entitled to say either.

Now, Perry's defense assumes the Mary is entitled to assert [M-1u]. Let's assume that Perry addresses the difficulties mentioned above concerning such a claim. Perry would now say that [M-1u] and [M-2u] have the same same subject matter content; and, therefore, that we can only identify the additional knowledge that [M-2u] expresses by examining the reflexive content.

Reflexive content imposes new truth conditions on Mary's statements. In this case, the reflexive content is that the experiential phenomenon "that is the origin of her old concept is the very one to which she is attending" (Perry, 2001, 148).

Assuming that "her old concept" referred to whatever concept she had associated with the term "tomato red" while still confined to her room, Perry's claim seems very dubious. How can the new experience be the origin of the pre-existing concept?

I don't deny that there is something like what Perry is calling reflexive content. One can easily make it more salient by imagining Mary speaking in the first person,

> [M-3u] I am now attending to *that* experiential phenomenon, tomato red.

In [M-3u], the target of the act of inner ostension is the experiential phenomenon with which Mary has become acquainted and to which she is attending as she speaks.

Now, [M-3u] is true iff the inner act of ostension is actually targeting an instance of the experiential phenomenon named tomato red.

[M-3u] could be false if Mary was deceived by a clever plant geneticist who

modified a tomato to look blue to people with ordinary color vision. Mary may have recognized the item she was given as a tomato based on shape, size or taste. She might then incorrectly think that she was seeing tomato red when in fact she was seeing a pale blue.

How does this discussion of the reflexive content of Mary's statement defend identity theory physicalism from the KA?

Intuitively, merely granting that there is reflexive content to certain statements doesn't undermine the KA. To have that rhetorical effect, one must also assert that there is no other knowledge gained. Perry makes this assertion indirectly by claiming that the subject matter content of [M-1u] and [M-2u] is the same.

An advocate of the KA need not accept this claim uncritically. It seems rather vulnerable to the charge that Perry has contrived an example by cherry picking a non-representative sample of from the pool of all statements Mary might make. Consider the following pair of statements.

> [M-4] I am proficient in using the term, "tomato red"; but, I am not acquainted with the experiential phenomenon to which that term refers.
>
> [M-5] I am proficient in using the term, "tomato red"; and, I am now acquainted with the experiential phenomenon to which that term refers.

I assume that it'll be uncontroversial that Mary could state [M-4] before her release and [M-5] after. I will also assume, although this may be somewhat more controversial, that [M-5] expresses the new knowledge that Mary gains from her experience.

Both of these statements are *mention* cases with respect to the term, "tomato red"; so, the use/mention distinction doesn't account for her new knowledge.

Both statements have reflexive content in that the inventory of experiential phenomena Mary is acquainted with must (in [M-5]) or must not (in [M-4]) include the referent of the term she mentions. However, only [M-5] incorporates her new knowledge.

However, it is not clear what grounds the antecedent physicalist might have for saying that the subject matter is the same with respect to the second clause in each statement. The subject matter of each statement concerns the items Mary is acquainted with. If we take an inventory of experiential phenomena with which Mary is acquainted before and after her first encounter with the tomato, we will find that she is acquainted with one more item after her encounter compared to before.

Thus, the truth conditions for [M-4] and [M-5] turn on the subject matter content. We now face the problem that Perry designed antecedent physicalism to avoid. A new subject matter, the new experiential phenomenon Mary has only now encountered, is a potential threat to physicalism.

Her knowledge is about something she didn't know – wasn't acquainted with – before.

We might make this (alleged) fact the subject of a statement Mary might make while she is gazing at the tomato.

> [M-6] I am now experiencing *that*, an experiential phenomenon I've never experienced before.

Now, there are no complications due to color terminology. The statement is true iff *that*, the experiential phenomenon to which she is attending as she makes the statement, is new to her. While the statement may be reflexive, it's clearly a statement where, in accordance with the subject matter assumption, we must be concerned about what Perry calls *the subject matter content*.

> The subject matter assumption may be put succinctly and almost, it seems, tautologically, as follows: The content of a belief is simply whatever is believed about whatever the belief is about. (Perry, 2001, 113)

Assuming that [M-6] expresses a belief at all (rather than, say, a phenomenal fact), the belief is that "I've never experienced *that* before". Clearly, it's about *that* experiential phenomenon.

Now, one might say that Mary (before her release) knew (by description) all about physical$_1$ phenomena related to color vision. After her release, she experiences a phenomenon she has never experienced before. At first glance, it certainly seems that there is more to know about than the physical$_1$ phenomena she already knew about.

However, since two ways of knowing are involved, knowing by description (knowing$_d$) and knowing by acquaintance (knowing$_a$), things are not so simple. The identity theorist might reply along the lines that Perry suggested earlier, with a claim something like the following Identity Hypothesis.

> **[IH-1] If the physical phenomenon that is known$_d$ (by description) is identical to the experiential phenomenon that is known$_a$ (by acquaintance), the answer to [DFQ-1] is "One".**

Now, presenting an argument in favor of the identity claim made in the antecedent of [IH-1] would involve arguing on behalf of some version of [KA:TNG-C2]. Leaving the antecedent unsupported while providing defensive arguments against some other possible conclusion creates the problem of the potentially vacuous conditional defense of physicalism.

### §3.3.3.1.2.2 In Search of an Informative Identity

As many identity theorists do, Perry assumes that the objective is to find an informative identity modeled after the ones Frege wrote about.

> With the knowledge argument and the modal arguments, it is helpful to put the debate in the context of Frege's problem about informative identities. It seems common sense that the reason a true thought of the form "A is B" might be informative, although "A is A" is not, is that the former involves two different ways of thinking of the same object; the information is simply that these are two ways of thinking of the same object. There can be two ways of thinking of properties and states, not only of things. I can think of the color of blood as "the color of blood" or as "red" or, while attending to a red object, as "this color." (Perry, 2001, 18-19)

I have no objection to an unprejudiced inquiry into whether physical/phenomenal identity claims are true or false; but, to be unprejudiced, such an inquiry could

not initially assuming either the identity or the non-identity of physical and experiential phenomena. Those are the conclusions toward one of which the epistemological journey is aimed.

To fit the Fregean model of an informative identity, the inquiry would need to start off with two distinct terms and proceed to show that the two terms refer to a single phenomenon (or object or property or whatever).

So let us imagine that Mary has a vast Concordance of Phenomenology and Terminology, CPT. The concordance lists all known physical phenomena such as neural firing patterns, NFP, that occur in the brain.[13] Each entry for an NFP consists of its designation, the prefix "NFP-" followed by a unique identifying number, plus a description in technical terms explaining its identifying characteristics. Some entries for NFP also list one or more terms for correlated or associated experiential phenomena.

Each entry for an experiential phenomena consists of a term in everyday language that an experiencing subject might use to designate an experiential phenomenon. It may also include references to any NFP or other physical phenomena that are known to be correlated with the occurrence of that experiential phenomenon.

Let us assume that Mary's CPT has an entry for tomato red which lists NFP-4738 as the correlated physical phenomenon; and, that it has an entry for NFP-4738 which lists tomato red as the name normal viewers give to the experiential phenomenon associated with that NFP.

As noted earlier, after her release and after her encounter with the tomato, Mary is able to say

**[M-2u] Tomato red is *that* experiential phenomenon**

and we would say that she recognizes tomato red.

Perry would say that this is the sort of informative identity claim that Frege identity theorists are searching for.

He uses different terminology, of course. Instead of speaking about experiential phenomena he speaks about subjective characters. Instead of using “tomato red” to designate that which she experiences, Perry imagines Mary's lessons using the term “$Q_R$”. Just as I imagine that Mary, before her release, is able to use “tomato red” without being acquainted with its referent, Perry imagines that Mary, before her release, is similarly able to use “$Q_R$”.

After her release, Perry imagines Mary saying something analogous to [M-2u]: “$Q_R$ is this subjective character”. This is the statement that he says is “… the version of Mary’s new knowledge that we’ll take as our official Frege problem”. (Perry, 2001, 100)

---

13 There is also space for listing physical phenomena that aren't NFP; and, Mary anticipates listing quantum microtubular computations as soon as their occurrence can be confirmed and their conditions of individuation become known. I will speak about NFPs only; but, nothing turns on whether the neural correlate of an experiential phenomenon is a NFP or something else as long as it is something of which Mary can have complete knowledge by description.

I disagree with Perry on this point. He hasn't introduced two terms; just one term and an opinion as to how it should be classified – as a subjective character rather than as an experiential phenomenon (or something else). That's analogous to saying “The Evening Star is that spot of light in the night sky” while pointing to a particular spot of light in the night sky. While possibly true, if the speaker is pointing to the correct spot of light, an informative identity claim of the sort we're looking for should be analogous to “The Evening Star is the Morning Star”, the claim Frege used to illustrate the problem of the informative identity.

In my view, one can only state a Fregean Identity Claim with two referring terms. Relevant to the present discussion the claim would be

> **[FIC-1] The physical phenomenon known$_d$ as NFP-4738 is identical to the experiential phenomenon known$_a$ as tomato red.**

If known to be true, this would be an informative identity claim. If known to be false, the following would be an informative non-identity claim.

> **[FNC-1] The physical phenomenon known$_d$ as NFP-4738 is not identical to the experiential phenomenon known$_a$ as tomato red.**

Perry makes no attempt to argue for [FIC-1]. He simply assumes that “$Q_R$” is both a term for the subjective character Mary experiences after her first encounter with the tomato and a term for something physical in the brain.

That's not good enough for a knowledge based response to Mary's plight.

If Mary invokes the grandmother clause of the revised gedanken, she will refuse to assume her way across an epistemological gap. Philosophers need to provide a principled basis for Mary to *conclude* rather than merely *assume* either that [FIC-1] or that [FNC-1] is true.

In the next subsection, I will present an argument for [FNC-1] resting on propositions that seem plausible to me. Following that, I will turn my attention to the question of what Mary learns upon her release with a view to deciding whether the knowledge she gains will support the claims that must be made to draw the conclusion that [FNC-1] is true.

Along the way, I'll consider whether identity theorists past and present provide any support for [FIC-1] or something analogous to it.

### §3.3.3.1.2.3 Argument for Physical/Experiential Non-Identity

In presenting a case for [KA:TNG-C3] or some more specific non-identity claim such as [FNC-1], the non-identity theorist takes direct aim at [IH-1] or any similar defense of physicalism based on a potentially vacuous conditional claim.

By [KA:TNG-1], Mary knows$_d$ all the facts her lessons can give her about physical phenomena; and, these facts are normally called *physical facts*; but, to control for conflation and equivocation, I will call them *physical$_1$ facts*. One might plausibly argue that whatever facts scientists discover about physical$_1$ phenomena can be written down and communicated to Mary; and, that Mary, being the brilliant student that she is, is able to learn those facts.

By [KA:TNG-2], Mary becomes acquainted with experiential phenomena she has

never experienced before, tomato red and other experiential color phenomena. She knows$_a$ the experiential phenomena which which she becomes acquainted as she experiences it.

Now, just having two terms, *experiential phenomena* and *physical phenomena*, doesn't guarantee that they are terms for distinct categories. The argument for physical/experiential non-identity, PEN, is an argument for distinctness.

> **[PEN-1] A physical$_1$ phenomenon occurs in an experiencer independent way.**

By this I mean that a physical$_1$ phenomenon occurs at a place, at a time and that its occurrence is not dependent on an experiencing subject.[14]

> **[PEN-2] An experiential phenomenon occurs in an experiencer dependent way.**

By this I mean that an experiential phenomenon occurs to an experiencing subject at a time; so, it is dependent for its occurrence on the existence of the subject.

> **[PEN-3] No phenomenon can occur in both an experiencer dependent way and in an experiencer independent way.**
>
> **[PEN-4] Hence, no phenomenon can be both a physical$_1$ phenomenon and an experiential phenomenon.**
>
> **[PEN-5] Hence, no experiential phenomenon is (identical to) a physical$_1$ phenomenon**
>
> **[PEN-6] If there are experiential phenomena they are not identical to physical$_1$ phenomena.**

This conclusion doesn't guarantee that there are any experiential phenomena at all; but, given that we have good grounds for rejecting eliminative materialism, there is something to talk about, the phenomena that can't be eliminated.

*But, how do we know that we are talking about phenomena in the first place?*

Given that one task facing neuroscientists and philosophers of consciousness is to explain the occurrence of first-person phenomenology, it seems only natural to say that the items of which first-person phenomenology consists are phenomena (appearances) rather than properties, objects or something else.

*How do we know that any of these claims are true?*

For purposes of the discussion of KA:TNG, then, the question is whether Mary's color experience provides any ground for holding that phenomenal colors (*tomato red* and/or *otamot green*) are experiential phenomena rather than physical$_1$ phenomena. The crucial point is whether Mary has any grounds for inferring or otherwise concluding that experiential phenomena occur in an

---

14 Some qualification with respect to subatomic phenomena may be needed to address the measurement problem of quantum mechanics; but, I don't think we need to worry about that here. It won't help the identity theory physicalist to claim that an immaterial causally efficacious consciousness must be postulated to explain the laws of physics.

experiencer dependent way.

*How will her experience tell Mary that [PEN-3] is true?*

I can make no sense of the negation of [PEN-3], the claim that there is a phenomenon that occurs in both an experiencer dependent and an experiencer independent way; so, I'm treating these as mutually exclusive categories.

Before her release, Mary $\text{knew}_{d}$ all about $\text{physical}_{1}$ phenomena such as NFP-4738 occurring in the brains of other people as they looked at ripe tomatoes. She knows that any number of people could measure a neural firing pattern in the brain of someone else.

There doesn't seem to be any basis for saying that when Mary finally experiences tomato red for herself, she gains the knowledge upon which to conclude that NFP-4738 is nothing more than an appearance to an experiencing subject, herself. It is, however, quite possible that she might gain the knowledge upon which to basis the conclusion that something is occurring to her in an experiencer dependent way; namely, *that* experiential phenomenon she's currently attending to.

*Is there a more formal version of this argument?*

(Note: Assume a universe of discourse consisting of all phenomena.)

```
[1]  (∀x)(Px → Nx)                          Assumption   {1}
```

For any x, if x is a $\text{physical}_{1}$ phenomenon (P), x occurs in an experiencer independent (N) way.

```
[2]  (∀x)(Qx → Dx)                          Assumption   {2}
```

For any x, if x is an experiential phenomenon (Q), x occurs in an experiencer dependent (D) way.

```
[3]  (∀x)¬(Dx & Nx)                         Assumption   {3}
```

No phenomenon can occur in both an experiencer dependent and an experiencer independent way.

```
[4]  (∃x)(Px & Qx)                          Assumption   {4}
```

There is something that is both a $\text{physical}_{1}$ phenomenon and an experiential phenomenon. This is the assumption introduced for the purposes of the *reductio ad absurdum*.

```
[5]  Pa & Qa                                4  Existential Instantiation   {4}
[6]  Pa                                     5  & Elimination   {4}
[7]  Qa                                     5  & Elimination   {4}
[8]  Pa → Na                                1  Universal Instantiation   {1}
[9]  Na                                     6,8  Modus Ponens   {1,4}
[10] Qa → Da                                2  Universal Instantiation   {2}
[11] Da                                     7,10  Modus Ponens   {2,4}
[12] Da & Na                                9,11  & Introduction  {1,2,4}
[13] ¬(Da & Na)                             3  Universal Instantiation   {3}
```

| | | |
|---|---|---|
| [14] (Da & Na) & -(Da & Na) | 12,13 & Introduction | {1,2,3,4} |
| [15] ¬(∃x)(Px & Qx) | 4,14 Reductio | {1,2,3} |
| [16] ¬(Pb & Qb) | 15 Existential Instantiation | {1,2,3} |
| [17] Qb | Assumption | {17} |
| [18] Pb | Assumption | {18} |
| [19] Pb & Qb | 17,18 & Introduction | {17,18} |
| [20] (Pb & Qb) & ¬(Pb & Qb) | 16,19 & Introduction | {1,2,3,17,18} |
| [21] ¬Pb | 18,20 Reductio | {1,2,3,17} |
| [22] Qb → ¬Pb | 17,21 Conditional Proof | {1,2,3} |
| [23] (∀x)(Qx → ¬Pb) | 22 Universal Instantiation | {1,2,3} |

For any phenomenon x, if it is an experiential phenomenon it is not a physical$_1$ phenomenon. Q.E.D.

### §3.3.3.2 What Does Mary Learn Upon Her Release?

The second premise of KA:TNG is both more explicit and more specific than Jackson's second premise. It is explicit as to the source Mary's new knowledge; and, it specifies what is known by acquaintance: experiential phenomena.

There are at least three ways to contest the second premise. One could deny that humans have the ability to know by acquaintance. I will counter this objection by defusing the supposed conflict between the acquaintance hypothesis and the main alternative to any acquaintance-based hypothesis, the ability hypothesis.

Alternatively, one could admit that humans have the ability to know by acquaintance but deny that, in becoming acquainted with phenomenal redness, Mary becomes acquainted with a phenomenon rather than a property or something else.

I will counter the second objection by showing (1) that acquaintance with experiential phenomena is the most natural way to understand of the notion of acquaintance; and, (2) that approaches to the KA based on an identity claim fail when we understand acquaintance as acquaintance with experiential phenomena.

Thirdly, even if one admits that, in experiencing, we become acquainted with experiential phenomena, one may deny that Mary learns anything that threatens identity theory physicalism. In reply, I will argue that what Mary learns is sufficient to ground an inference that an experiential phenomenon *is only as* it appears; and, further, that this inference supports the conclusion reached deductively via the PEN argument.

#### §3.3.3.2.1 The Additional Abilities Hypothesis

The Ability Hypothesis (hereafter AH) developed by Nemirow and popularized by Lewis (1988) is that knowing what it is like to experience some quality, Q, is nothing more than a set of abilities, the so-called Lewis abilities: the abilities to remember experiencing Q, to imagine experiencing Q and to recognize Q when

experiencing it again.

The weakness of AH is that Nemirow seems opposed to explaining how someone might acquire (or extend) these abilities on the basis of phenomenal information. He asks us to consider a proposition such as Mary might assert when she first encounters a tomato; for example,

> (X) This is what it's like to see red.
>
> … a proponent of KA might charge AH with begging the question against KA by assuming that the Lewis abilities alone, without phenomenal information, provide the cognitive background required for the understanding of (X). According to this charge, knowledge of phenomenal information explains the acquisition of the Lewis abilities, and therefore the Lewis abilities may not be invoked to explain away the understanding that underlies such knowledge. However, this defense reveals a weakness in the very position that it tries to secure. Although KA starts out as an argument against physicalism, in its current incarnation it is reduced to relying upon knowledge of phenomenal information to "explain" the presence of the Lewis abilities. If this "explanation" were known to be correct at the outset, there would be no need for KA in order to establish the existence of phenomenal information in the first place. (Nemirow, 2007, 45-46)

This defense misrepresents the intent of the KA. It *recognizes* the existence of phenomenal information, information about experience. What it attempts to establish is that phenomenal information is not physical information. Few deny that Mary can report information about her experiences; and, Nemirow is not among them. He's actually rather liberal in acknowledging that

> AH is designed to deal only with certain types of knowledge claims by reducing them to assertions about practical abilities. It asserts that statements about knowing what an experience is like are statements about abilities rather than claims to propositional knowledge. In short, AH is a theory about a certain kind of knowing, and nothing more. Defenders of AH may freely acknowledge that certain terms (such as "this taste") refer to experiences. No part of AH compels its defenders to deny that there are experiences; that we have a vocabulary to refer to them; or that they may be referenced by demonstratives. Moreover, as a theory about a certain type of knowing, AH is not committed to the view that our naked references to experiences are reducible to statements about abilities. (Still to be considered, however, is the import for KA and AH of these acknowledged limitations on the ambitions of AH.) (Nemirow, 2007, 38-39)

If we have an a vocabulary for referring to experience and we use it, are we not exercising our *ability* to refer to experience?

Apparently, where I differ from Nemirow is that I would say that humans have relevant abilities in addition to the Lewis abilities. Specifically, I would say that humans also have at least three other relevant abilities:

1. The ability to notice their experiences and the experiential phenomena that make up their experience;
2. The ability to refer to their experiences and to the experiential phenomena that make up their experience; and,
3. The ability to comment on their experience by generating statements about experience generally or about the experiential phenomena that make up

some particular experience.

So, in my view, when Mary first encounters the tomato, she notices phenomenal redness and she generates a statement about her experience which she reports to her audience of data-hungry experimental philosophers. Perhaps, she initially exclaims “So, *that* is tomato red” and, after a few moments of reflection, reports “I am now experiencing tomato red” while the experience is ongoing.

To distinguish my perspective from one in which *only* the Lewis abilities are considered relevant to a discussion of the KA, I call my view the Additional Abilities Hypothesis, AAH.

This recognition of additional relevant human abilities provides the basis for a theoretical unification of two responses to the KA that are generally considered to be in conflict, the ability hypothesis and the acquaintance hypothesis. After all is said and done, what does an acquaintance hypothesis need besides the ability to notice, refer to and comment on experience? In my view, those abilities constitute the ability to become acquainted with … whatever one becomes acquainted with. The only further element an acquaintance hypothesis needs to be useful is a claim about the nature of that with which one becomes acquainted.

In my view, Mary becomes acquainted with experiential *phenomena*.

Thus, in contrast to Nemirow, becoming acquainted with experiential phenomena underwrites both Mary's acquisition of phenomenal information and her abilities to imagine, remember and recognize experiential phenomena.

### §3.3.3.2.2 Instantiation in Experience

Terrence Horgan admits that Mary learns something when she first encounters her tomato; and, claims that her new knowledge or new information can only be expressed by an indexical proposition such as

> [IP-1] Seeing ripe tomatoes has *this* property

where it is understood that

> ... 'this property' is used to designate the colour-*quale* that is instantiated in her present experience. (We shall call this property *phenomenal redness*. It should not be confused, of course, with the redness-property instantiated in the tomatoes themselves.) (Horgan, 1984b, 151)

Since I deny that tomatoes have a redness property, I don't think I'll have any trouble distinguishing that alleged property from phenomenal redness; but, the truth value of color realism is beside the point that Horgan is making – that phenomenal redness is a *property* (rather than the phenomenon I take it to be). That difference aside, I quite agree that phenomenal redness is instantiated in Mary's experience while she is looking at the tomato; and, that Mary acquires new knowledge or new information when she first sees the tomato.

*Does admitting that Mary has information about a color quale or knowledge of phenomenal redness threaten identity theory physicalism?*

Horgan denies that the new information Mary acquires poses any threat to physicalism because physicalists may claim that the new information is

*ontologically physical information*; meaning, information that only refers to physical entities despite being expressed in language that does not explicitly refer to any physical entity.

> The information is new not because the *quale* she experiences is a non-physical property, but because she is now acquainted with this property from the experiential perspective. (Horgan, 1984b, 151)

Horgan is making an identity claim: that the quale, phenomenal redness, Mary experiences from the experiential perspective is identical to a property that she already knew about via some other perspective; presumably, her knowledge by scientific description.

If one *assumes* that the phenomenal redness Mary becomes acquainted with is identical to a physical property that she already knew about; then, one might argue that, in stating [IP-1], Mary only referred to physical entities she already knew about; but, I'm not willing to assume that Mary will simply assume that she has become acquainted with a property rather than with a phenomenon.

In short, Horgan has taken rhetorical liberties with Mary's philosophical position by assuming in [IP-1] that she will take phenomenal redness to be a property. I *could* put different words into Mary's mouth by imagining her refusing to assert [IP-1] in favor of

[IP-2] That which I am seeing is an experiential phenomenon, tomato red

but, I won't as that would create an impasse. Horgan and I could each rationalize our belief that Mary should interpret her experience in the terminology of our own theory; and, that will get us nowhere.

So, instead, I'll imagine that Mary wants to state her new knowledge without expressing any pre-existing philosophical intuitions she may have (from her reflections on the phenomenology associated with other senses); and, that she is aloof as to the debate between Horgan and myself. In this scenario Mary simply asserts

[IP-3] So, *that* is tomato red

and that, after some philosophical reflection, she tweets the Friends of Mary network,

[IP-4] I am now experiencing tomato red!

Both [IP-3] and [IP-4] express Mary's new knowledge but leave it open for debate just what Mary becomes acquainted with. The point of rejecting both [IP-1] and [IP-2] in favor of [IP-3] or [IP-4] is that advocates of the KA don't have to exaggerate the extent of the knowledge Mary gains from her immediately given sensory experience by suggesting that a philosophical judgment (that phenomenal redness is a physical property or that it is an experiential phenomenon or whatever) is itself given in experience. It takes considerable philosophical reflection effort to form a judgment as to whether phenomenal redness falls under the concept of a phenomenon or a property.

Clearly, Horgan and I agree that, in experiencing, Mary becomes acquainted with something instantiated in her experience; but, we make different claims about

what sort of thing that something is. He would say that *tomato red* is a phenomenal property; whereas, I would say that *tomato red* is an experiential phenomenon.

*Do we have alternate vocabularies or competing philosophical intuitions?*

In my view, “property” and “phenomenon” are not synonymous terms; and, consequently, debates concerning knowledge arguments will turn on whether the language of discourse classifies *tomato red* as a property or a phenomenon.

This will become clear as we consider whether an argument parallel to the one Horgan makes in terms of properties but expressed in terms of phenomena will succeed or fail in the circumstances at hand, where we limit Mary's knowledge claim to [IP-3] or [IP-4].

Horgan's argument is that Mary's knowledge claims do not threaten physicalism because Mary is expressing ontologically physical information; meaning that Mary is referring to a physical entity of some sort, despite not using any language that explicitly refers to a physical entity.

From Mary's point of view, she uses *tomato red* to refer to the particular shade of phenomenal redness she the experiences while staring at the tomato.

*How do we know that Mary is referring to a physical entity whenever she refers to tomato red, a name for a particular shade of phenomenal redness?*

The only relation between *tomato red* and whatever brain activity is correlated with experiencing *tomato red* that would guarantee that a reference to *tomato red* is a reference to that brain activity *whether Mary intended it as such or not* is the identity relation.

Horgan explains why in the course of deflecting a faulty argument he imagines an advocate of the KA giving in opposition to the claim that Mary may be held to have referred to a brain state merely because she referred to *tomato red* – whether she wants to be so held or not.[15]

> Perhaps it will be replied that the phrase 'this property' in [IP-1] cannot designate a physical property, because if it did then [IP-1] would express a piece of information which Mary had already: viz., the information that ripe-tomato perceptions possess the given physical property. But this reply ignores the all-important intensionality of the notion of information. Even though Superman is Clark Kent, nevertheless we must distinguish between the information that Superman can fly and the information that Clark Kent can fly. (Horgan, 1984b, 151-152)

It's true that Lois Lane does not know that Clark Kent can fly; but, it doesn't follow that Lois Lane knows that Clark Kent can't fly. The absence of evidence

---

15 Mary's situation is reminiscent of the plight suffered by Otto, Dennett's foil in Consciousness Explained. Otto is depicted as a qualiaphile who tries to refer to the pink ring he experiences in a neon color spreading illusion only to be told (rather imperiously, in my view) that he, Otto, is mistaken because he, Dennett, has identified “'the way it is with me' with the sum total of all the idiosyncratic reactive dispositions inherent in my nervous system as a result of my being confronted by a certain pattern of stimulation” (1991, 386-387). In consequence of this identification, Dennett tells Otto, “When you say 'This is my quale', what you are singling out, or referring to, whether you realize it or not, is your idiosyncratic complex of dispositions” (1991, 389). As we shall see, however, Otto has a reply which is available to Mary as well.

(for non-identity) is not evidence of absence (of identity).

Horgan is not arguing *for* identity theory; he does not present an affirmative defense showing that the identity claim is true. He is arguing *from* identity theory. He assumes a physical/phenomenal identity and argues that it is defensible; meaning, that knowledge-based arguments against the identity claim are not sufficient to show that the assumed identity is false.

The non-identity theorist may reply that the defense has a weakness: it assumes that the argument against which it is defending is cast is terms of phenomenal *properties*; specifically, in the case at hand, the assumption is that phenomenal redness is a phenomenal property.[16]

Now, there are any number of philosophers whose knowledge-based arguments against physicalism are based on the assumption that phenomenal redness is a phenomenal property; so, I am not faulting defenders of the identity theory for defending their position against such arguments. I am, however, faulting defenders of identity theory for trying to perpetrate a rhetorical analogue of the French military strategy that led to the construction of the Maginot Line.

Even assuming that the French Maginot Line would have provided France an excellent defense against a German attack *on the Maginot Line*, the French defensive strategy was flawed because nothing compelled the Germans to attack the Maginot Line head on; and, as is well-known, they simply went around it.

Similarly, nothing compels us to assume that Mary becomes acquainted with a phenomenal property or a qualitative property or a property of some other kind when she first becomes acquainted with phenomenal redness.

Let us now consider Mary's situation in KA:TNG whose second premise assumes that, in experiencing, we become acquainted with an experiential *phenomenon* rather than a property of some kind. I would say that *tomato red*, the particular shade of phenomenal redness that Mary experiences – that becomes instantiated in her experience – when she first encounters a ripe tomato, is the experiential phenomenon that Mary becomes acquainted with when she sees a ripe tomato.

So far, these claims parallel Horgan's claims; so, the question is whether an attack on the identity claim parallel to the one Horgan deflects with “the Clark Kent is Superman” example can be similarly deflected.

Suppose I say (or imagine Mary, the phenomenalist, saying)

> [MP-1] I know$_d$ that NFP-4738 is a physical phenomenon detectable by my cerebroscope.
>
> [MP-2] I now know$_a$ that *tomato red* is *that* experiential phenomenon.
>
> [MP-3] I wonder whether NFP-4738 is identical to *tomato red*.

Could it be the case that the physical phenomenon, NFP-4738, is identical to the experiential phenomenon, *tomato red*? Could it be the case that *tomato red* just

---

16 Nothing in the argument that follows turns whether tomato red is considered a phenomenal property rather than a qualitative property. The contrast being made is the contrast between property and phenomenon.

*is* NFP-4738 despite the undisputed fact that I don't know it *as* NFP-4738?

Taking NFP-4738 and *tomato red* as properties, Horgan holds that it could still be the case that NFP-4738 and *tomato red* are identical.

Taking NFP-4738 and *tomato red* as phenomena allows us to access what we already know about phenomena as such; for example, that a phenomenon is an appearance distinct from any physical reality of which it may be an appearance. This additional knowledge justifies the conclusion that NFP-4738 and *tomato red* are not identical.

Our notion of phenomena as appearance also includes the notion that there is some ontology that accounts for a given phenomenon. Consequently, it is entirely reasonable to invoke some physical reality in attempting to account for the appearance to the experiencing subject of some experiential phenomenon. But, its status as an appearance precludes the possibility that an experiential phenomenon is identical to the physical reality of which it is merely an appearance.

Consequently, the identity claim fails; and, it is possible to refer to tomato red, the experiential phenomenon without referring to any neural phenomenon with which it may be associated.

And that is what I claim I am doing when I say “the Crescent Moon is not identical to the Full Moon”. I am using “Crescent Moon” to refer to a crescent shaped patch of bright light in the night sky and “Full Moon” to refer to a circular shaped patch of bright light in the night sky. I can accept that both appearances are appearances of Earth's large natural satellite, Luna, the Moon, the self-identical physical reality of which both appearances are distinct appearances. I commit no self-contradiction in saying that the Crescent Moon is not identical to the Full Moon despite the fact that both are appearances of the same underlying physical reality; because, in referring to the appearance I am not thereby referring to the physical reality to which the appearance is not identical.

Now, it is not possible for Horgan or other identity theory physicalists to reply, as they did under the assumption that tomato red is a property rather than a phenomenon, that the phenomenon Mary became acquainted with from the experiential perspective is the same phenomenon she already new about from a different perspective.

A phenomenon is an appearance. An experiential phenomenon is as it appears to its experiencer via the experiential perspective. I know about the physical phenomenon, NFP-4738, via another perspective, the perspective of inferences made from scientific observation.

*Is NFP-4738 another appearance?*

If so, NFP-4738 and tomato red are distinct appearances. Even if they are each an appearance of the same *tertium quid* they'd be distinct from each other just as the Crescent Moon is distinct from the Full Moon. As appearances, neither NFP-4738 and nor tomato red would be identical to that *tertium quid* of which each is a distinct appearance.

*Is NFP-4738 the physical reality behind the appearance known as tomato red?*

If so, tomato red is not identical to the physical reality of which it is merely an appearance.

*Is NFP-4738 the cause of the appearance known as tomato red?*

If so, tomato red is not identical to its own cause.

*Is the appearance known$_a$ as tomato red identical to the appearance known$_d$ as NFP-4738?*

If tomato red and NFP-4738 are identical, then something that occurs in an experiencer dependent way would be identical to something that occurs in an experiencer independent way. But, that's a contradiction.

#### §3.3.3.2.2.1 Evaluating Horgan's Defensive Strategy

In the paper under discussion, Horgan (1984b) assumes the identity of physical and phenomenal properties; and, undertakes to show that the identity theory physicalist has a defensive reply to various attacks a non-identity theorist might make. Horgan's rhetorical strategy seems to be that, when tomato red is taken to be a phenomenal property, non-identity theorists can't claim to know that tomato red is not identical to a physical property they may not know anything about.

Clearly, the success of this strategy depends on a tactical move, taking tomato red to be a property rather than a phenomenon; but, the non-identity theorist is not required play along. The anti-identity argument can easily be recast in terms of phenomena rather than properties. When this is done, the Horgan defense fails because the non-identity theorist *can* say that a physical phenomenon I know about via scientific description is thereby known as something that exists in an experiencer independent way, not as an experiential phenomenon, an appearance-to-me that I am directly acquainted with or immediately aware of and which occurs in an experiencer dependent way.

Indeed, with the added assumption that a phenomenon is instantiated by its occurrence, we can say that experiential phenomena exist in an experiencer dependent way and that physical$_1$ phenomena exist in an experiencer independent way.

In essence, Horgan's rhetorical strategy seems modeled after French military strategy. He needs to argue that a quale of experience *can't* be taken as a phenomenon rather than as a property. As it stands, he simply assumes that non-identity theorists *won't* ever do that.

And yet, here I am doing just that. I affirm the existence of first-person phenomenology; and, I take the phenomenalist perspective – that phenomenal redness and other items of interest to a study of first-person phenomenology are … phenomena.

#### §3.3.3.2.2.2 Evaluating Horgan's Identity Claim

Horgan presents his own identity theory elsewhere; but, as we shall see, the identity claim is an assumption rather than a conclusion.

> I propose a two-part theory. Phenomenal state-types (i.e., qualia) are to be identified with neurophysiological state-types; the identities involved are necessary identities, because the qualia-names involved (as well as the neurophysiological state-names) are rigid designators. Non-phenomenal state-types, on the other hand, are to be construed functionally, in accordance with either first-order or second-order functionalism, whichever proves more viable. (Horgan, 1984a, 460)

This sounds like it would certainly qualify as a relevant identity theory; meaning, one which, if true, might have some bearing on Mary's epistemological journey. However, Horgan's theory seems, at first glance, anyway, to have a dualistic element. According to Horgan, we can't assume that all mental state-types are either phenomenal or non-phenomenal but not both.

> It may turn out that many garden-variety mental state-types are really hybrid types, involving both a phenomenal component and a non-phenomenal component. A plausible candidate for such hybrid status, I suggest, is the philosopher's favorite mental state: pain. I think there are really two state-types instantiated in any clear-cut instance of pain: (1) *phenomenal* pain, the 'raw feel' of pain experiences (and the element not definable functionally); and (2) *functional* pain, the state-type which, by definition, has typical causes such as harmful forces impinging upon the creature's surface, and typical effects such as avoidance-behavior. Hybrid types, on the view I am proposing, are instantiated when both the relevant purely-phenomenal type and the relevant purely-functional type are instantiated -- even if these latter are not explicitly countenanced in everyday mentalese. (Horgan, 1984a, 460-461)

What saves Horgan's theory from the suggestion of dualism is a footnote claiming that, under both the first-order and second-order versions of his theory, “the term 'phenomenal pain' denotes brain state B” (1984a, 461). Under the first-order version of the theory, “functional pain” also refers to brain state B, albeit nonrigidly; but, it is not clear how having two terms that both refer to the same brain state means that two state-types are instantiated.

There are other difficulties. For example, it is not clear what it means to say that experiential phenomena are not functionally definable *and* they are identical to physical phenomena which, presumably, are functionally definable. Would it mean that, after scientists finish cataloging all the physical phenomena that constitutes every brain function, there will be some unexplained brain activity that escapes their inventory?

In any case, Horgan doesn't actually argue that the claimed identity of the physical and the phenomenal is true; only that (1) it has pragmatic value for philosophers who want to defeat arguments for non-identity claims; and, (2) it is itself defensible in that the identity theorist can defeat objections to the identity claim. Horgan is particularly interested in a defense to the potentially dualistic conclusions of Kripkean arguments.

Horgan points out that Kripke asks us to imagine someone whose brain exhibits the physical phenomena usually associate with experiencing pain but who is experiencing a tickle sensation instead. Horgan complains that, from the ability to imagine such a situation, Kripkeans conclude that such a situation actually is possible which, given standard possible worlds thinking, means that in some possible world the physical phenomena associated with pain is not co-instantiated with the experiential phenomenon of pain. It follows that the

experiential phenomenon of pain (pain-Q) is not identical to the physical phenomena associated with it (call it physical$_1$ pain or functional pain or pain-P).

Horgan admits that "this argument must be turned aside, if the theory I propose is to be viable". However, to turn aside the Kripkean argument, Horgan must argue that one can't actually imagine Mr Tickle; and, so he does. But, he must assume his own identity theory as a basis for the claim that we can't imagine Mr. Tickle.

> I have argued that if the proposed identity theory is correct, then our seeming ability to imagine Mr. Tickle can be explained away as a psychological illusion which rests upon a subtle fallacy: the as/is fallacy, as I've called it. (Horgan, 1984a, 467)

Clearly, if his theory is correct everyone who disagrees with it is wrong; and, all arguments against his theory would have to be fallacious; but, whatever its pragmatic value as a defense against arguments for non-identity claims, as an argument for the truth of the identity theory itself, Horgan's argument would be circular.

The main selling point of Horgan's identity theory is its rhetorical function: it provides a defense to one particular family of arguments for the theory of non-identity. Someone in the market for a defense to arguments for the theory of physical/experiential non-identity may simply assume a theory of physical/experiential identity; arguments to the contrary would have to be fallacious because they contradict the identity claim.

Unfortunately, we have no reason independent of identity theory to believe that the ability to imagine Mr. Tickle is illusionary. Moreover, a recent scientific discovery, human echolocation, seems to show that such a scenario is not only imaginable, it is actual.

It has been known for some time that some blind humans can learn about the objects in their vicinity by generating sounds and listening to the echoes. However, it has recently come to light that blind human echolocators process information from their echoes in the visual cortex (Thaler et al., 2011). While the full story of human echolocation has not yet been told, it seems clear that I can imagine someone in a certain brain state with activity in the visual cortex and imagine that person as having an auditory experience rather than a visual experience.

*Are we at an impasse?*

I have no doubt that competing philosophical intuitions are in play. For some the intuition of identity is as strong as the intuition of distinctness is for others; but, that doesn't necessarily mean that every dispute will degenerate into an impasse of conflicting question-begging arguments.

The non-identity theorist offers an argument against the identity of physical and experiential phenomena based on the existence conditions for each. For the experiential phenomenon of redness that I experience after inducing an afterimage to be identical to some physical phenomenon, something that exists in an experiencer dependent way would have to be identical to something that exists in an experiencer independent way (and vice versa).

Horgan hasn't offered an argument for the truth of the identity claim itself. He has merely offered it for its pragmatic value. For example, he points out that it is the only way to preserve a naturalistic view of humans and the causal closure principle without making experiential phenomena epiphenomenal. That is not an argument for the truth of the identity claim. It is an argument for the compatibility of the identity claim with other claims commonly made by physicalists and materialists.

The closest Horgan comes to defending the truth of his identity claim comes in the closing paragraph. He lists the various replies that he has available against the following objection.

> … it might be objected that phenomenal painfulness cannot be a physico-chemical property, for the following reason. Since 'phenomenally painful' differs in meaning from any physico-chemical predicate, it must have a sense that is distinct from the sense of any physico-chemical predicate. Hence there must be some *property* that constitutes the sense of 'phenomenally painful', and that does not constitute the sense of any physico-chemical predicate. (Horgan, 1984a, 469)

Among the replies available is this one:

> … there is at least one fairly popular identity-criterion for properties which is compatible with the contention that phenomenal properties are identical with physico-chemical properties: viz., that two properties are identical if and only if they are *necessarily coextensive*. This criterion is not easily satisfied, but I contend that it indeed is satisfied in the case of phenomenal and physico-chemical properties – notwithstanding the fact that the relevant identity-statements will be empirical rather than *a priori*. (Horgan, 1984a, 469)

Now Horgan does not actually produce the evidence and/or argument that *shows* that this criterion for property identity is indeed satisfied; so, the argument as it stands is incomplete.

Even worse is that Horgan has again produced an argument that turns on whether we consider tomato red (or some other instance of phenomenality) as a phenomenon or as a property. If tomato red is a phenomenon rather than a property, that criterion is simply not applicable.

*Is there an analogous criterion for phenomenon identity?* I argue that there is no such analogous criterion.

In view of Horgan's use of 'instantiated' in reference to qualia (1984a), one may reasonably take 'coextensive' to mean 'co-instantiated'. If that is the case then the identity via co-instantiation argument will fail. The half-moon and the Moon are co-instantiated; but, they are not identical. The half-moon is an appearance and as such is not identical to the physical reality of which it is an appearance.

Of course, an identity theorist could reply that the full moon and the Moon are not always co-instantiated. When the crescent moon is instantiated, the Moon is instantiated but the full moon is not.

The failure to be co-instantiated is conclusive proof that the two items are not identical; but, what if the two are always co-instantiated. Suppose it turns out that the experiential phenomenon of pain is always instantiated when certain physical$_1$ phenomena, C-fibers firing, are instantiated. That doesn't necessarily

make them identical. Co-instantiation doesn't rule out the possibility that firing C-fibers are the physical cause of a phenomenal effect. Furthermore, co-instantiation does not rule out the possibility that phenomenal pain is the appearance to an experiencing subject of an underlying physical reality to which that appearance is not identical.

The identity theorist whose evidence is co-instantiation must rule out the cause/effect, the appearance/reality and any other non-identity relation when arguing that identity is the only relation that explains co-instantiation.

Horgan made no attempt to do so.

### §3.3.3.2.2.3 Crossing the Rubicon

I've been concerned since §3.3.3.2 with defending the second premise of KA:TNG – that in experiencing we are acquainted with phenomena rather than properties.

*What has been accomplished?*

I've examined two papers by Terence Horgan, one which presented an identity theory and one which presented a reply to Jackson's KA based on the assumption that a relevant identity theory is true.[17]

I chose to focus on Horgan's work because, despite being an identity theory, it *sounds like* phenomenon dualism in one key respect. Horgan speaks of hybrid types requiring the co-instantiation of a phenomenal state-type and a function state-type. Although I prefer speaking of phenomena rather than state-types, I have no objection to the assumption that scientists will continue to describe brain activity in terms of its functions. Philosophers of consciousness *should* build on that premise.

Horgan's claim that a term like *phenomenal redness* refers to a brain state can only be true if the experiential phenomenon that makes a phenomenal state-type the state-type it is and the physical phenomenon that makes a functional state-type the state-type it is are numerically identical, one and the same phenomenon. Eliminate that assumption and the result is a version of phenomenon dualism.

Horgan speaks of a color quale being instantiated in Mary's experience; and, I think that is quite right. This position is in direct opposition to Dennett's eliminative approach to the afterimage. He says that it is part of our experience; but, that it doesn't exist. In my view, a phenomenon is instantiated by its occurrence; so, an afterimage exists – is instantiated in experience – while it is occurring. Consequently, I agree with Horgan in opposition to Dennett.

I agree with Horgan that the color quale instantiated in Mary's experience when she first sees a tomato may be called *phenomenal redness*; although, I generally prefer to speak about *tomato red*, particular shade of phenomenal redness experienced when looking at a ripe tomato. We disagree as to whether phenomenal redness is a property or a phenomenon; and, we disagree as to

---

17 In my view, a claim of physical/phenomenal identity is relevant to KA:TNG. Theories proposing other identities would not be relevant. I will return to this point shortly, arguing that some popular identity theories are not relevant to KA:TNG.

whether phenomenal redness is identical to something physical.

It may be objected that choosing to classify phenomenal redness as a property or as a phenomenon is a matter of linguistic style; but, it's not like choosing to say that Mary encounters a to*may*to rather than a to*mah*to. The choice between taking *tomato red* to be a property and taking *tomato red* to be a phenomenon has consequences because the language of discourse is not neutral on this point.[18]

Rejecting Dennett's eliminative approach to first-person phenomenology is like crossing the Rubicon. There *will* be consequences.

Affirming that there is first-person phenomenology raises the question of how to characterize items such as phenomenal redness that would be included in an inventory of first-person phenomenology and that must be explained by a theory of experience. Are these items properties or phenomena?

Intuitively, the most plausible answer is that those items are phenomena; and, I will call those who say that the items that make up first-person phenomenology are phenomena, *phenomenists*. I will call those who say that the items that make up first-person phenomenology are themselves properties, *propertyists*.

I am a phenomenist.

Ostensibly, both phenomenists and propertyists affirm the existence of first-person phenomenology; but, if items such as phenomenal redness are properties rather than phenomena, it is difficult to see what is leftover that might be a first-person phenomenon; so, classifying items such as phenomenal redness as properties rather than phenomena may well turn out to be a disguised form of eliminativism as to first-person phenomena. The elimination of first person phenomena would certainly explain why the propertyist has no trouble defending his or her identification of phenomenal redness (taken as a phenomenal property) and some third-person property.

By classifying phenomenal redness as a phenomenon rather than a property, the phenomenist preserves the ability to argue that the relation between an experiential phenomenon and a physical phenomenon is the appearance/reality relation – a non-identity relation. This makes it more difficult for the identity theorist to argue for a physical/phenomenal identity. The identity theorist would have to rule out the appearance/reality relation as well as all other non-identity relations.

Suppose that an identity theory physicalist wanted to adapt the Horgan defense to oppose knowledge-based arguments that assume that phenomenal redness is an experiential phenomenon rather than a property. What then?

Could the identity theory physicalist develop criteria for phenomenal identity analogous to the criteria for property identity that Horgon mentions and argue on the basis of those criteria for the identity of physical and experiential phenomena?

---

18 Ironically, despite the controversy surrounding the use of “quale(s)/qualia”, that term is agnostic as to whether a quale is a property or a phenomenon. Nevertheless, at some point one must make a choice as to which concept qualia fall under, property or phenomenon.

That remains to be seen; but, I am skeptical of the prospects for showing that a measurable, objectively detectable physical phenomenon is nothing other than an appearance to an experiencing subject.

### §3.3.3.3 Is the Nagel/Churchland Argument Repairable?

As noted previously, Churchland glimpsed the question around which the debate swirls; and, it seems clear that his answer to [DFQ-1] is "One".

Curiously, Churchland not provide an affirmative defense of phenomenon monism. He argues that advocates of phenomenon dualism fail to demonstrate the existence of an existential phenomenon not identical to some physical phenomenon occurring in the brain; but, he makes no attempt to argue for the identity claim, leaving that as something scientists will have to discover.

Eventually, we will look elsewhere for those who offer affirmative defenses of an identity theory. At the moment, I want to confront Churchland's negative defense, his argument that the non-identity of experiential and physical phenomena has not been demonstrated.

Churchland objects to arguments for non-identity based on what is known about experiential phenomena by introspecting, which I will take to be the equivalent of knowing by acquaintance.

After criticizing one of Thomas Nagel's arguments based on considerations relying on the intensionality of knowledge, Churchland reports having received from Nagel an updated version of the argument, as follows (but presented here my own identifiers for each step).

> [NC-1] My mental states are knowable by me by introspection.
>
> [NC-2] My brain states are not knowable by me by introspection.
>
> [NC-3] (therefore) My mental states ≠ my brain states

Churchland acknowledges that this argument is valid; and, in particular, that it is free of the intensional fallacy he found in the original version of Nagel's argument; but, ...

> But now the reductionist is in a position to insist that the argument contains a false premise: premise 2. At the very least, he can insist that (2) begs the question. For if mental states are indeed identical with brain states, then it is really brain states that we have been introspecting all along, though without appreciating their fine-grained nature. (Churchland, 1985, 21)

#### §3.3.3.3.1 Premise 2: The Repairable Flaw

There is a flaw associated with premise 2; but, it appears in Churchland's statement in opposition to premise 2: “if mental states are indeed identical with brain states, then it is really brain states that we have been introspecting all along”.

In my view, introspecting is nothing other than knowing by acquaintance[19]; so, I

---

19 In principle, Churchland is free to adopt a view that distinguishes knowing by introspecting

can introspect only what I'm acquainted with, what I'm immediately aware of.

Suppose that someone someday somehow proves that a brain state, something with an objective mode of existence, is identical to the color of an afterimage, something with a subjective mode of existence, an appearance to an experiencing subject. *Even then*, it will not be the case that I am directly or immediately aware of my brain states. *Why not?* Because, as Churchland reminds us, knowing is an intensional context.

Just as Lois Lane can know that Superman can fly without knowing that Clark Kent can fly, a claim that I am acquainted with the experiential phenomena that constitutes my experience does not support a claim that I am acquainted with the physical phenomena that constitute my brain states – irregardless of opinions as to the relation between physical and experiential phenomena.

For his objection to Nagel's reformulated premise 2 to fly as stated, Churchland would need to explain why knowing by introspecting is not an intensional context, something that doesn't seem very likely.

There is another possibility as well. I suspect that what Churchland is trying to say is that we might discover that the mental state one introspects is identical to a brain state one learns about some other way. If that is the case, then his objection to premise 2 contains the same flaw that exists in premise 1.

#### §3.3.3.3.2 Premise 1: The Fatal Ambiguity

In my view, [NC-1] is fatally flawed because “mental state” is inherently ambiguous.

*What object is my mental state a state of?*

Suppose we were talking about an immaterial mind of the Cartesian sort. Clearly, the states of such a mind would be mental states; and, equally clearly, they would not be brain states. [NC-3] follows directly from the way mental states are defined in [NC-1]; so, [NC-2] is irrelevant.

Suppose that a mental state is the state of a mental object other than a Cartesian-style mind; perhaps, an object such as was postulated by early versions of the sense-data theory, an object whose properties determined our experience. Again, [NC-3] follows directly from the way mental states are defined in [NC-1]; so, again, [NC-2] is irrelevant.

Physicalists would certainly favor a less inflammatory interpretation: taking a mental state to be a physical state of the brain which, for some reason, is *also* a mental state of the brain. But, then the mental states of my brain are obviously brain states. This time the conclusion is false; but, once again, [NC-2] is irrelevant.

I believe the Nagel/Churchland argument[20] is repairable; but, we must first

---

from knowing by acquaintance and allows him to know by introspecting something he isn't immediately aware of via introspection; but,

20 So-called because, while formulated by Nagel, it was blessed (declared formally valid) by

clarify the meaning of *a mental state of a brain*.

*What makes a brain state a mental state of a brain?*

In my view, a brain is in a certain state when it instantiates or exhibits a physical phenomenon, such as a neural firing pattern, that determines *which* brain state that brain is in. When some phenomenality, some experiential phenomenon, is associated with that brain state, I would consider that brain state to *also* be a mental state of the brain.

Now, if an identity theorist is able to show that the experiential phenomenon that makes a brain state a mental state of a brain is identical to the physical phenomenon that makes that brain state the physical state that it is, he or she could reasonably say that the experiential phenomenon in question is nothing other than a physical phenomenon of the brain. In that case, the non-identity theory will have been falsified; and, dualism will have been averted.

Conversely, if a non-identity theorist is able to show that the experiential phenomenon that makes a brain state a mental state of a brain is not identical to the physical phenomenon that makes that brain state the physical state that it is, he or she could reasonably say that the experiential phenomenon in question is something other than a physical phenomenon of the brain. Identity theory will have been falsified; and, the door will be opened to the claim that dualism has been achieved (or lapsed into, as the case may be).

In my view, the question turns on whether experiential phenomena have a subjective mode of existence. If so, they can't be identical to physical phenomena which have an objective mode of existence.

### §3.3.3.3.3 Repairing the Nagel/Churchland Argument

The Nagel/Churchland argument as repaired by Polanik:

> [NCP-1] Experiential phenomena occurring to me are knowable by me by acquaintance.
>
> [NCP-2] Physical phenomena occurring in my brain are not knowable by me by acquaintance.
>
> [NCP-3] (therefore) Experiential phenomena ≠ physical phenomena.

This argument, NCP, is free of the ambiguities introduced by referring to brain states and mental states; and, retains the formally validity of its predecessor.

*Is NCP immune to the objection Churchland raised against its predecessor?*

Now, it would seem that, if some phenomenon, P, is both a physical phenomenon and an experiential phenomenon, it would be knowable by me by acquaintance and not knowable by me by acquaintance, a contradiction; hence, [NCP-3]. But, let us imagine that Churchland renews his objection.

Taking *tomato red* as our target item, it would be agreed by all that, in knowing *tomato red* by acquaintance, I don't know it *as* a physical phenomenon. The

---

Churchland.

essence of Churchland's objection is that, despite not being known as a physical phenomenon, tomato red might nevertheless *be* a physical phenomenon.

However, the non-identity theorist argues that, in looking at a tomato, there is other knowledge to consider besides our knowledge by acquaintance with an experiential phenomenon such as *tomato red*. We have at least the following as well:

1. Our knowledge about knowing by acquaintance; and,
2. Our knowledge by description of physical phenomena;

#### §3.3.3.3.3.1 Knowledge about Knowing by Acquaintance

A person can live a full life without being reflectively aware that he or she knows experiential phenomena by acquaintance. But the philosopher, having affirmed the existence of acquaintance knowledge, will likely reflect on the nature of knowing by acquaintance and develop a theory about it so that it may then play a role in ongoing philosophical disputes.

For present purposes, I'll adopt the view of knowing by acquaintance developed by Balog, who considers her account to be neutral as between physicalistic and dualistic accounts of qualia and argues that a successful account of phenomenal concepts will explain acquaintance and other epistemic puzzles about consciousness.

> We know our conscious states not by inference but by immediate acquaintance which gives us direct, unmediated, substantial insight into their nature. (Balog, 2012, 21)

Such a view of knowing by acquaintance incorporates what has come to be called the Thesis of Revelation. As formulated by Johnston (1992b, 223), Revelation is the thesis that “The intrinsic nature of canary yellow is fully revealed by a standard visual experience as of a canary yellow thing” and similarly for other colors and other experiential phenomena.

Henceforth, I will take the Thesis of Revelation as this principle:

> **[TR] An experiential phenomenon *is* only as it appears.**

What is revealed in experience is just what used to be considered given in experience, in this case, the instance of canary yellow. The classification of canary yellow as an experiential phenomenon rather than a phenomenal property or something else is neither given nor revealed in experience. That comes as conceptual knowledge; once we reject the eliminative approach to first-person phenomenology, we must make a judgment as to whether we're dealing with phenomena or properties or whatever.

It seems that Balog would agree with this much at least. Although she usually speaks of phenomenality in terms of phenomenal states and phenomenal experiences, she seems to accept that all such items are in the overall category of *phenomenon*. In her view, physical accounts of consciousness seem inadequate because

> ... there is a strong tendency to think that an insight into the nature of a phenomenon (e.g., via acquaintance) should lead one a priori to any other insights into the nature of the same phenomenon (e.g., via neuro-scientific analysis)... (Balog, 2012, 31)

Equally clearly, Balog assumes that it is possible for a phenomenon with which we are acquainted can be the identical to a phenomenon of which we have scientific knowledge. However, she does not explain how that is possible; so, here we part company.

In my view, once we know that we are dealing with a phenomenon, we also know analytically that a we are dealing with an appearance.

Now, if the Thesis of Revelation is true, an experiential phenomenon *is* ***only as it appears*** *to its experiencer*; so, there is no identity of physical and experiential phenomena. It is not the case that an objectively detectable or measurable physical phenomenon is nothing other than an appearance to an experiencing subject.

#### §3.3.3.3.3.2 Knowledge by Description of Physical Phenomena

In knowing a physical phenomenon, NFP-4738, by description I know it *as* a physical phenomenon; and, I also know that I don't know it *as* an experiential phenomenon. However, I reject the possibility that NFP-4738 may nevertheless *be* nothing other than an experiential phenomenon, an appearance to a subject. From the fact of knowing NFP-4378 by scientific description, I conclude that it exists objectively and not subjectively as an experiential phenomenon.

Consequently, *tomato red* is *only as* it appears, an appearance, a phenomenon.

#### §3.3.3.3.4 Further Considerations

These considerations seems sufficient to defend [NCP-3] against Churchland's objection; but, if more is needed, it's available. I refer the reader to the defenses raised in §3.3.3.2.2 against the claim that what is known via description can be identical to what is known by acquaintance.

Finally, I point out that advocates of the theory that something that exists subjectively may somehow be identical to something that exists objectively have yet to explain how I might come to know this alleged identity.

#### §3.3.3.3.4.1 How Would I Come to Know the Alleged Identity?

If I don't know from immediate experience, with which I am (directly) acquainted, that I am acquainted with a brain state, how would I find out?

I suppose I could find out the way that Mary might find out, by looking it up in her Concordance of Phenomenology and Terminology, CPT, a vast catalog cross-referencing neural phenomena and color terminology.

Suppose that Mary, upon her release, is presented with a large selection of unlabeled color swatches (all the same size and shape) and is asked to pick out tomato red. It's doubtful that she could do it without further information. According to Knut Norby, a totally achromatic color vision scientist, Mary would

not be able to do it.

> Mary will not know what to call the various color sensations unless she makes use of noncolor information; for example, knowing that a rose is red, she may recognize the form of a rose and deduce that therefore its color must be red. Achromatic people do this all the time by memorizing such facts as that fire engines are red, violets are blue, grass is green, lemons are yellow, and so on. (Nordby, 2007, 79)

Okay; so, let's give Mary some further information.

Suppose we give Mary a talking neuroscope with which she may examine her own brain. She looks up “tomato red” in her Concordance of Phenomenology and Terminology and discovers that *tomato red* is the name of the color phenomenon experienced by someone whose brain is exhibiting neural firing pattern, NFP-4738. Now Mary connects her neuroscope to her own brain, looks at the color swatches one at a time and waits for her neuroscope to specify the name of the NFP it detects. When it says “I have detected NFP-4738”, Mary concludes that she is experiencing tomato red as a result of looking at that particular unlabeled color swatch; and, she utters the expected remark, “Ah! So, *that* is tomato red”.

Let us assume that Mary could do this with complete accuracy. Looking at the color swatch that provoked the neuroscope to report detecting NFP-4738 counts as being being directly acquainted with the experiential phenomenon, *tomato red*; but, I deny that this process counts as being *directly* acquainted with a brain state.

*Would this process count as indirect acquaintance with a brain state?*

In my view, knowing by acquaintance presupposes a sense of directness or immediacy. “I am acquainted with experiential phenomena” means the same as “I am directly acquainted with experiential phenomena”. Consequently, Mary's conclusion that she is experiencing *tomato red* is not an instance of knowing by acquaintance.

*What process supports Mary's conclusion that she is experiencing tomato red?*

I call the process by which Mary infers that she is looking at something that is tomato red a *forensic inference* because it is the sort of inference that forensic experts display on TV crime scene investigative dramas. It presupposes expert knowledge. For example, suppose I come across a corpse while hiking through the forest. I might notice that it is being eaten by beetles; but, I can't infer anything useful from that fact; so, I call the police. They bring along a forensic expert who knows from extensive lessons (and, possibly, some personal research) that, given the longitude, latitude and elevation of the scene, the body would have to have been dumped in the forest on or about a specified date to become infested with that species of beetle.

In the case at hand, Mary does not know by direct acquaintance that she is undergoing brain state, NFP-4738. She draws that conclusion by relying on her talking neuroscope.

When she first sees the color swatch she becomes acquainted with *tomato red*; but, doesn't know it by that name until her neuroscope reports detecting brain state NFP-4738. At that point, she may rely on the expert knowledge embodied in

her Concordance of Phenomenology and Terminology to support the forensic inference that what she is seeing is known as *tomato red*.

I suppose that someone could arbitrarily classify what I am calling a forensic inference as a variety of acquaintance; but, that seems forced. In my view listening to the neuroscope should not count as direct acquaintance nor as introspection, to use Churchland's terminology.

Churchland would concede that Mary is using two ways of knowing; but, he holds out hope for an identity claim.

> In short, the difference between a person who knows all about the visual cortex but has never enjoyed a sensation of red, and a person who knows no neuroscience but knows well the sensation of red, may reside not in what is respectively known by each (brain states by the former, qualia by the latter), but rather in the different type of knowledge each has of exactly the same thing. The difference is in the manner of the knowing, not in the nature of the thing(s) known. (Churchland, 1985, 24)

However, Churchland does not explain how Mary is supposed to derive the conclusion that NFP-4738 which she knew about from her lessons is identical to *tomato red* which she only knows about by experiencing it.

Until we know that, there is no basis for the claim that the two ways of knowing provide knowledge of exactly the same thing. Instead, there is every reason to conclude that knowledge by acquaintance provides knowledge of experiential phenomena; and, that knowledge by description provides knowledge of physical phenomena. And, in that case, there is every reason to believe that one may infer the non-identity of experiential and physical phenomena from their existence conditions.

If we assume that a phenomenon is instantiated by its occurrence, we can see that each phenomenon is instantiated under different conditions. Experiential phenomena are instantiated by their occurrence *to an experiencer* at a time. Physical phenomena are instantiated simply by their occurrence at a time and place. Experiential phenomena have a subjective, experiencer dependent mode of existence; whereas, physical phenomena have an objective, experiencer independent mode of existence.

They can't be identical.

### §3.3.3.3.4.2 Accounting for Belief in the Possibility of Identity

If it is not possible for experiential phenomena and physical phenomena to be identical, what could account for belief that philosophers will someday demonstrate the identity or that scientists will someday show the world that identity theorists were right all along?

In claiming knowledge of experiential phenomena by acquaintance, my knowledge is *of* the phenomena of which I am immediately aware.

Now, suppose that, by some process of forensic inference, I conclude that my knowledge is *of* something that I believe accounts for the experiential phenomena in question -- a neural firing pattern, complex quantum computations in the microtubules of the brain, information transformation during wave

function collapse or whatever.

Is my knowledge knowledge of the same thing in each case? No. I've equivocated between knowledge of an appearance (the experiential phenomenon, the direct or immediate object of my awareness) and knowledge of a physical reality which may account for that appearance.

A phenomenon is an appearance. It may be an appearance of something else, the reality behind the appearance; but, while an appearance is a *phenomenological* reality to its experiencer, it is not identical to the *physical* reality of which it is merely an appearance. For example, the Half Moon is an appearance of the Moon. The Full Moon is an appearance of the Moon. Neither appearance is identical to the Moon itself because, if they were, they'd be identical to each other by the transitivity of identity; but, they aren't identical to each other. Not only are they numerically distinct because they occur at different times, they are distinguishable. They look different. I can tell them apart.

Hence, I can say without contradicting the law of identity both that the Moon is self-identical; and, that the Full Moon is not identical to the Half Moon.

Clearly, I may also say “The Full Moon is the Moon” and “The Half Moon is the Moon” and similarly for other phases of the Moon; but, in such statements, the “is” can't be the is of identity; otherwise, the transitivity of identity would lead us into contradiction.

*What sense of “is” is in play here?*

### §3.3.3.3.4.3 Is There an Is of Appearance?

It may be time to recognize another use of “is” – the *is of appearance* – distinct from the is of identity and the other recognized uses of “is”.

In saying something like “The Full Moon is the Moon” I'm not saying that the Full Moon is (identical to) the Moon. The appearance (a phenomenal reality) is not identical to the (physical) reality of which it is merely an appearance. I'm saying that the Full Moon is (an appearance of) the Moon.

There is also a reversed usage. Advertisements for films sometimes say things like “Sean Connery is James Bond”. Clearly, the claim is not that the actor is identical to the character he plays. The claim is that Sean Connery is *appearing as* James Bond in the film being advertised.[21]

Similarly, one can reverse “the Full Moon is (an appearance of) the Moon”. We'd get “the Moon is (appears as) the Full Moon”. This sort of statement would be entirely appropriate if, for example, a child were to ask me “Daddy, what is the 'Gibbous Moon'”? I could truthfully reply, “The Gibbous Moon is (the appearance

---

21 As I recall, such statements were common in advertising for re-releases of the earlier Bond movies after Roger Moore began appearing in the role of James Bond; and, particularly when the 'is' in the statement is typographically emphasized, may reflect the writer's belief that Sean Connery was the best of the many actors who played that role. Sean Connery *is* James Bond! See “Picking The Best Bond” at http://www.npr.org/sections/monkeysee/2012/10/05/162175114/picking-the-best-bond-connery-and-craig-rise-to-the-top

of) the Moon when it looks almost full”.

Recognizing the is of appearance would make it less difficult to explain why “the tomato is red” means “the tomato appears red to me”. In abandoning a naive realism as to color we learn that redness is an element of phenomenal reality (appearance) not an element of physical reality. One then continues to affirm the statement in contention (in the debate between color realists and irrealists); but, one has a different understanding of it.[22]

### §3.3.3.4 Instructively Irrelevant Identity Claims

While there are philosophers (e.g. Papineau, 2002) who admit that materialism requires affirming counterintuitive physical/phenomenal identities, some so-called identity theorists are actually eliminative as to experiential phenomena.

For example, J. J. C. Smart distinguishes between an afterimage and the experience of having an afterimage; and, clearly states, "I am not arguing that the after-image is a brain-process, but that the experience of having an after-image is a brain-process” (Smart, 1959, 150). Smart turns eliminative as to the afterimage itself. “There is, in a sense, no such thing as an after-image or a sense-datum, though there is such a thing as the experience of having an image” (Smart, 1959, 151).

Clearly, from Smart's perspective, no brain process is an afterimage; and, the non-identity theorist agrees.

*What is at issue, then?*

Fundamentally, the *existence* of experiential phenomena is at issue; but, the existential/eliminative dispute seems to manifest itself in disputes over the language of discourse; particularly, *the* meaning of “experience”.[23]

Following Smart, identity theorists who are eliminative as to experiential phenomena designate some brain process they admit is not an afterimage as an “experience”. In contrast, the non-identity theorist – particularly, the phenomenon dualist – would say that the afterimage itself is the experience. Both agree that no brain process is an afterimage; but, only the non-identity theorist concludes that an experiential phenomenon such as an afterimage is not a physical phenomenon such as would constitute a brain process. The eliminative identity theorist simply denies the existence of experiential phenomena.

Clearly, then, eliminative identity theorists and non-identity theorists use “experience” differently.

I do not propose to decide which use is the one and only, true and correct use of

---

22 Similarly, in the debate over whether water is identical to $H_2O$ or merely constituted by it, both sides affirm the claim “Water is $H_2O$“ but differ as to whether to take it as using the is of identity or the is of constitution. Consequently, a philosopher could switch sides in the debate and continue to affirm the statement in contention; but, understand it differently.

23 Sadly, it appears that, despite heroic efforts devoted to discussing the so-called Hard Problem of experience, philosophers of consciousness do not agree as to the referent of “experience” and related widely-used terms. (I'm shocked! Shocked, I say!)

“experience”. Instead, I will merely distinguish these uses in the following way. I will use “experience-Q” in place of the non-identity theorists' use of “experience” to refer to experiential phenomena not identical to any physical phenomena. I will use “experience-P” in place of the traditional/eliminative identity theorists' use of “experience” to designate certain brain processes.

Here the suffixes “P” and “Q” allude to Chalmer's abstract description of conceivability arguments as the conjunction, P & -Q, where P is the “conjunction of all microphysical truths about the universe” and Q is “an arbitrary phenomenal truth about the universe” (2010, 107).

The items that Smart claims to designate by “experience”, brain-processes, would all be included within the microphysical description of the physical universe. These items constitute experience-P.

The items that the non-identity theorist claims to designate by “experience” are claimed to be Q-facts. These items, if there are any such items, constitute experience-Q.

*How does this difference relate to KA:TNG?*.

Upon her release, Mary sees a ripe tomato and experiences an phenomenal redness for the first time. After gazing at her tomato for a long time, she looks away, toward a white wall, and sees a greenish afterimage. According to the eliminative identity theorist, the afterimage does not itself exist; yet, somehow, Mary has her first experience of phenomenal greenness.

The *eliminative identity theorist* has the task of explaining how Mary can learn something new, what phenomenal greenness is like, by experiencing an afterimage that does not exist.

The *existential identity theorist* has the task of explaining how Mary can learn something new, what phenomenal greenness is like, by experiencing an afterimage that exists but is identical to something she already knows about in some other way.

Both the existential and the eliminative identity theorist agree that the correct answer to [DFQ-1] is that there is only one fundamental kind of phenomenon; but, their argumentative tasks are different.

The tasks facing the eliminative identity theorist are:

1. To show that there are no experiential phenomena at all; and,
2. To show either that Mary doesn't learn anything from her new experiences or that what she learns doesn't threaten eliminative materialism.

The tasks facing the existential identity theorist are:

1. To show that the experiential phenomena instantiated in Mary's experience, phenomenal redness and phenomenal greenness, are each identical to some physical phenomenon occurring in her brain; and,
2. To show either that Mary doesn't learn anything from her new experiences or that what she learns doesn't threaten identity theory physicalism.

Intuitively, Mary learns something from seeing her first tomato and something

else from seeing her first afterimage. This is the challenge of KA:TNG itself; and, the second item on each task list responds to it.

* * *

In the next few subsections, I will review some well known identity theories with a view to deciding whether they can shed any light on Mary's plight.

### §3.3.3.4.1 David K Lewis

Many philosophers distinguish experiencing construed as an act, state or event from the intentional objects of such an act, state or event. Lewis, for example, theorizes that "every experience is identical with some physical state. Specifically, with some neurochemical state" (Lewis, 1966, 17). However, to forestall confusion as to what he is attempting to identify with neurochemical states, he offers this guide to interpreting what "experience" refers to.

> We must understand that the identity theory asserts that certain physical states are experiences, introspectible processes or activities, not that they are the supposed intentional objects that experiences are experiences of. If these objects of experience really exist separate from experiences of them, or even as abstract parts thereof, they may well also be something physical. Perhaps they are also neural, or perhaps they are abstract constituents of veridically perceived surroundings, or perhaps they are something else, or nothing at all; but that is another story. So I am not claiming that an experience of seeing red, say, is itself somehow a red neural state. (Lewis, 1966, 18)

I, too, distinguish the act of experiencing from its object; *but*, I use “experience” and related terms such as “experiential phenomenon” to refer to the intentional object of my experiencing rather than to the neural activity that occurs while I am experiencing whatever I am experiencing.

In my view, the intentional object of my experiencing is my experience; meaning, it is a discrete part of the seamless whole of my experience that I single out for further commentary. Any such discrete bit or *aspect* of experience is an experiential phenomenon, an item to be included in a comprehensive inventory of first-person phenomenology.

With respect to KA:TNG, I would say that the token of phenomenal redness that Mary experiences is an experiential phenomenon; that it exists while it is occurring; and, that it is not identical to any physical phenomenon.

*Is Lewis' version of identity theory of any help to those defending identity theory physicalism from these claims?*

Clearly, Lewis is not trying to show that the intentional objects of experiencing are nothing other than neurochemical states; he freely admits he is talking about ... something else; so, he is not contesting my claim that experiential phenomena are not identical to physical phenomena.

Equally clearly, Lewis avoids taking a stand as to the existence or non-existence of the intentional objects of experiencing; so, he is not – at least, not *here* where he states his identity theory – contesting my claim that an experiential phenomenon exists while it is occurring to its experiencer.

Elsewhere, when discussing the KA, Lewis turns eliminative as to experiential

phenomena. “... suppose the phenomenal aspect of the world had been absent altogether, as we materialists think it is”. (Lewis, 1988, 95)

Thus, while Lewis' views may help those defending phenomenon monism from KA:TNG, it is not because those views state a relevant *identity* claim. It is because they state a relevant *eliminative* claim.

### §3.3.3.4.2 Gilbert Harman

Harman sets himself the task of showing that functionalism can account for the subjective feel of experience, what it is like “to undergo this or that experience” (Harman, 1990, 33). So, his task is analogous to that of the identity theorist who is trying to show that identity theory can account for same subjective feel of experience.

Harman's argument is based on “distinguishing properties of the object of experience from properties of the experience of an object” (Harman, 1990, 31). As an example, he asks us to consider “the experience of having a pain in your right leg”.

> It is very tempting to confuse features of what you experience as happening in your leg with intrinsic features of your experience. But the happening in your leg that you are presented with is the intentional object of your experience; it is not the experience itself. The content of your experience is that there is a disturbance of a certain specific sort in your right leg. The intentional object of the experience is an event located in your right leg. The experience itself is not located in your right leg. If the experience is anywhere specific, it is somewhere in your brain. (Harman, 1990, 40)

I agree that, in the event of suffering an injury to my right leg, I am presented with a pain I take to be in my right leg. That experiential phenomenon (or pain quale in more traditional language) is what Harman calls the *intentional object* of an experience. I say that the experiential phenomenon *is* my experience. Harman says it is not.

That Harman satisfies himself that experience as he defines it is accounted for by functionalism is beside the point. What he calls experience is the neural correlate of the experiential phenomenon in question; and, yes, that neural correlate is in the brain.

To show that functionalism leaves nothing unaccounted for, Harman must show either

1. That experience as I define it does not exist in any sense; or,
2. That experience as I define it is accounted for by functionalism.

Harman makes no attempt argue for the former. He sidesteps it by saying

> I am quite willing to believe that there are not really any nonexistent objects and that apparent talk of such objects should be analyzed away somehow. I do not see that it is my job to resolve this issue. (Harman, 1990, 37)

Harman's argument only shows that functionalism can, in his view, account for what he calls experience (or what I'd call experiencing-P); but, that's irrelevant. At issue is experiencing-Q, the experiential phenomena that constitute my

experience (in my view) or that constitute the intentional objects of experiencing (in Harman's view). Without them, we would be zombies rather than humans. Virtually everyone concedes that functionalism can account for the brain activity that accompanies reports of experiencing-Q. The difference between humans and zombies (if such are possible at all) is that, in humans, experiencing-Q accompanies the brain activity that constitutes experiencing-P (the nervous system reactivity/functionality that we undergo in response to incoming stimuli).

It doesn't really matter how scientists explain brain activity. I concede (as I hope any reasonable thinker would concede) that scientists will eventually provide both a complete description of brain activity in terms of the functions it performs and a complete explanation of how those functions are performed in terms of lower level processes possibly invoking quantum phenomena.

Harman is trying to say that nothing is left out of the functionalist's account but he tells us precisely what he leaves out. Harman also adopts a highly contentious linguistic constraint.

> When we say the pain or afterimage exists, we mean that the experience exists. When we say that the afterimage is on the wall or that the pain is in your leg, we are talking about the location of the intentional object of that experience. (Harman, 1990, 40)

When I say that an afterimage I am experiencing exists, I mean that *the experience as I define it* exists; meaning, the experiential phenomena, the circular patch of greenness constituting my afterimage, exists. When Harman says the same thing he means something completely different because he has defined “experience” so it refers to something other than experiential phenomena. He means that some physical phenomenon exists in the brain.

Harman's argument is that we are not normally aware of the intrinsic features of experience *as he defines it*; so, nothing we are normally aware of is left out of the functionalist's account. It is relatively easy to image a scenario in which we could be aware of what Harman considers an intrinsic feature of experience.

> Suppose David undergoes brain surgery which he watches in a mirror. Suppose that he sees certain intrinsic features of the firing of certain neurons in his brain and suppose that the firing of these neurons is the realization of part of the experience he is having at that moment. In that case, David is aware of intrinsic features of his experience. But that way of being aware of intrinsic features of experience is not incompatible with functionalism. (Harman, 1990, 41)

But, normally, one is not aware of what Harman calls the intrinsic features of experience (what I'd call experiencing-P); and, few philosophers claim otherwise. Although Churchland speculates that someday we might learn to directly introspect our brain states, we clearly can't do that now.

Harman considers a claim that a qualiaphile might make: in experiencing, one is aware of the intrinsic qualities of (what I call) experience; namely, the experiential phenomena I am actually aware of. For examples, I am aware of the phenomenal redness I experience while looking at a tomato and phenomenal greenness when looking at the afterimage of a tomato.

> My reply to this objection is that it cannot be defended without confusing intrinsic features of the intentional object of experience with intrinsic features of the experience.

> Apart from that confusion, there is no reason to think that we are ever aware of the relevant intrinsic features of our experiences. (Harman, 1990, 42)

In my view, the experiential phenomena that Harman calls the intentional objects of experience *are* the experience they constitute; and, yes, that's precisely how the objection is defended.

Certainly, a part of this debate concerns the proper definition(s) of “experience”; but, either language can be used to pick out what the non-identity theorist wants explained: the intentional objects of experience in Harman's language or the experiential phenomena in my language. Naturally, the question *I* would ask is: *how does the brain generate experiential phenomena?* And, I would argue that, so far, the functionalists haven't specified the neural function responsible for generating experiential phenomena not identical to any physical phenomena.

Of course, by asking the question that interests me, I may be begging the question against the identity theorist. Clearly, Harman is free to beg an alternate question. He might assume that experiential phenomena (as I use that term) are identical to physical phenomena; hence, there is no brain function that generates experiential phenomena not identical to physical phenomena; hence, nothing is left out of the functionalist's account of experiencing. However, he has not done so; and, having distinguished the experiential phenomena that I call experience from the brain activity that he calls experience, it's not likely he can do so without self-contradiction.

Alternately, Harman could simply deny the existence of experiential phenomena; but, he hasn't yet done that either.

Consequently, experiential phenomena escape the functionalist account of brain activity; at least, as Harman tells the tale. He fails to support either the eliminative or the identity outcomes to KA:TNG.

### §3.3.3.4.3 Byrne and Hilbert

In the course of defending color realism, Byrne & Hilbert (2003, 5) tell us that when “someone looks at a tomato in good light, she undergoes a visual experience”. The language of *undergoing an experience* strongly suggests that Byrne & Hilbert are referring to experiencing-P, some nervous system reaction to incoming stimuli. Sure enough, a short time later they cite Harman for the need to avoid conflating “the properties of an experience with the properties represented by the experience” and conclude that “An experience of a tomato is an event, *presumably a neural event of some kind*” (my emphasis).

Byrne & Hilbert are assuming that the neurological events that occur while I am experiencing are my experience; whereas, I am assuming that the experiential phenomena that occur while I am experiencing a tomato are my experience. We have competing identity claims as to what experience is. I am saying that my experience is nothing other than the experiential phenomena I experience. Byrne & Hilbert are saying that their experience is nothing other than the neurological phenomena that occur while they are experiencing.

Byrne & Hilbert favor the language of events over the language of phenomena;

so, to translate into their terminology, I'm willing to say that the occurrence of an experiential phenomenon, phenomenal redness, say, is an experiential event. I'm also willing to assume that, while an experiential event is ongoing (while an experiential phenomenon is occurring to its experiencer), physical events (the occurrence of physical phenomena) in the brain of the experiencer may be relevant to whatever account of the experiential phenomenon is provided; and, that such physical events may be called *neural events*.

If Byrne & Hilbert want to *show*, by evidence and argument, that an experiential event is identical to a physical event, they are free to give it their best shot. However, I am not willing to simply assume from the getgo that an experiential event is identical to the physical event with which it is associated.

Neither am I willing to assume that experiential events simply do not exist; although, ultimately, Byrne & Hilbert seem to have taken the eliminative option.

Byrne & Hilbert appear to dismiss the intentional object of the neural event when it is not a physical object. They say that when looking at a tomato, the tomato is the intentional object of the experience and stress that it is not itself an event because it does not occur. This would appear to exclude afterimages from being the intentional objects of experience; afterimages occur.

*Does an afterimage exist while it is occurring?*

> When one has an experience of a red circular afterimage, the content of the experience is – to a first approximation – that there is a red circular patch at a certain location. But this proposition is simply false. There is no red circular patch – not even in some internal mental realm. (Byrne & Hilbert, 2003, 5)

Clearly, then, Byrne & Hilbert turn eliminative when it comes to afterimages and, by extension, experiential phenomena generally. In my view, if we deny the existence of the red circular patch and all color patches of other colors and shapes we have denied the existence of color experience.

In contrast, Byrne & Hilbert claim that some neural events are experiences; that part of the content of such a neural event is a proposition; and, in the case of looking at a tomato, *the* proposition contained in that neural event is “that the tomato is red”.

> … it might visually appear to a subject that there is a red bulgy object on the table. The proposition that there is a red bulgy object on the table is part of the content of the subject’s experience. (Byrne & Hilbert, 2003, 5)

I may be old school; but, I expect scientists rather than philosophers to investigate neural events. Should *scientists* ever find a proposition in whatever neural event accompanies an experiential event such as the occurrence to an experiencing I of an instance of phenomenal redness, I will be so informed. Until then, I'm just not buying into the claim that philosophers sitting in their armchairs can detect the propositions contained in neural events that are invisible to the naked eye.

This puts me in conflict with a key claim of representationalism, that neural events have propositional content. I am eliminativist as to the propositional content of experience; but, this is not the place to critique representationalism in

detail. Instead, I merely offer an alternate thesis based on AAH, the Additional Abilities Hypothesis: a subject of experience has the ability to comment on the experiential phenomena of which it is aware. If my commentary contains a proposition, I will claim to have generated that proposition and I will deny having discovered it lurking in a neural event I'm not even aware of.

Suppose I look at a tomato under optimal viewing conditions and generate one of the following propositions.

1. That tomato is red.
2. That tomato looks red to me.
3. That tomato looks red to me; therefore, that tomato is red.
4. That tomato looks red to me; but, I know it isn't actually red at all.

Proposition 1 clearly supports color realism. It is the proposition Byrne and Hilbert find in the neural event they assume is the referent of “my experience of the tomato”. Proposition 4 clearly supports color irrealism. Proposition 2 might be considered agnostic as to color realism; but, it is also incomplete in the sense that it could be expanded. Someone might say that, fully expanded, Proposition 2 becomes Proposition 3; but, someone else might say that it becomes Proposition 4 instead.

In any case, there is no empirical evidence that Proposition 1 is the one and only, true and correct proposition to be found in the neural event that allegedly *is* the experience of seeing a tomato.

Thus, one obvious advantage the Additional Abilities Hypothesis offers is that the experiencer can generate a wider range of comments than the one proposition favored by representationalists generally and color realists in particular. Another advantage is that one need not postulate the color properties unknown to color scientists which color realists must postulate.[24]

When Mary is released from her room, we might imagine her exclaiming, “Ahh! So, that is *tomato red*” when her gaze falls upon a ripe tomato or a blooming rose of an appropriate variety. It is difficult to imagine her concluding that phenomenal color is something other that what she sees when she sees her first tomato. Yet this is the claim that Byrne & Hilbert make. They then propose to argue for the identification of phenomenal color and physical color.

They take physical color to be color properties of ordinary objects such as tomatoes that are identical to ordinary physical properties, for example, surface reflectance properties already known to scientists; but, we can hardly overlook the fact that these two sets of properties, those known to color scientists and those postulated by philosophers, function differently in representationalism.

According to the representationalist, my experience-P (brain activity) represents

---

24 Chalmers is accused of property dualism for postulating the existence of properties unknown to physical scientists to explain experience; but, physicalists who deny color realism look the other way when color realists postulate color properties unknown to physical scientists to explain color experience. I reject this inconsistency. In my view, color realism is an instance of property dualism.

that the tomato I see is red. My experience does not represent that the tomato reflects light of certain wavelengths. But the color realist says that phenomenal color is identical to physical color. According to representationalists generally and color realists in particular, my experience contains the proposition that the tomato is red not the proposition that tomatoes reflect light of certain wavelengths. Surely my experience of looking at a tomato doesn't contain the proposition “That tomato reflects light in the 650 nm range”. But that physical phenomenon is what is actually explained by the absorption/reflectance properties of the surface of the tomato.

Byrne and Hilbert briefly recite the argument from illusion.

> Consider the experience of a red circular afterimage, produced by fixating on a green circular patch for a minute or so, and then looking at a white wall. It is perennially tempting to suppose that there is something red and circular that the perceiver is aware of. If there is, then because there is nothing red and circular in the world external to the perceiver, there must be something red and circular in the perceiver’s internal world – something mental, presumably, since nothing in the brain is red and circular. This red circular thing is a *sense datum*. (Byrne & Hilbert, 2003, 5)

They then say that they will assume that arguments against the existence of sense data are successful. Their representationalism is a response.

While there are many forms of the sense data theory, I only want to avoid the extreme forms of it. Consequently, I have no objection to Byrne and Hilbert's assumption that arguments against sense data in the strong, Russellian sense were successful. However, I want to retain the assumption that something – an experiential phenomenon, to be specific – is instantiated in the experiencer's stream of experiences; so, I'm not willing to simply assume that arguments effective against Russellian sense data are automatically equally effective against weaker sense data theories.

The issue will be whether a theory weak enough to deflect the criticisms leveled against stronger sense data theories is still strong enough to survive eliminativist accounts and escape the accounts offered by identity theory physicalists.

### §3.3.3.4.4 U. T. Place and The Identity Crisis of Identity Theory

According to Tim Crane (1995), physicalists who are not eliminative as to the mental hold that the mental is physical. However, ...

> In earlier physicalist literature, the ‘is’ in the phrase ‘the mental is physical’ was understood as the ‘is’ of strict identity. But recently physicalists have tended to understand the ‘is’ as something closer to the ‘is’ of constitution. To say that everything is physical in this sense is to say that everything either is a physical entity *or* is constituted by or composed of physical entities. This kind of physicalism admits that there are non-physical things — but they are exhaustively constituted by, or composed of, physical things. (Crane, 1995, 212)

Now, from the beginning, U.T. Place (1956) stated his "identity" theory in terms of composition; and, J.J.C Smart contends that "Place spoke of constitution rather than of identity" (Smart, 2007, §2).

For present purposes, I will assume that a theory of composition is a theory of

constitution. There are those who disagree, of course.[25] But, nothing in my argument turns on the outcome of that debate.

I am more concerned with the appearance of equivocation in Place's work as to whether his theory of composition is a theory of identity or a theory of non-identity.

In favor of the conclusion that Place' theory of composition is a theory of non-identity is his admission that identity is a symmetrical relation whereas composition is not. If A is identical to B, B is identical to A. However, if A type things are constituted or composed of B type things, B type things are not constituted or composed of A type things.

However, Place undermines that conclusion by suggesting that, despite having two descriptions, there is only one item being described.

> Nevertheless, although there is still an element of asymmetry between the macroscopic and the microscopic description whereby the microscopic explains the macroscopic and not vice versa, this is not the sort of asymmetry which is incompatible with asserting the symmetrical relationship of identity as far as the common referent of the two descriptions is concerned. (Place and Schneider, 2013)

Logically, identity is self-identity, a = a. Every thing is itself rather than something else. If two descriptions have a common referent, we have a true identity theory. There is only one item being described and that item is necessarily self-identical no matter how many descriptions of it we have; so, the formula, a = a, is satisfied.

On the other hand, when Place describes the relation he is actually talking about, he says “... the proposed identity relation between consciousness and the brain process or complex of brain processes in which it consists” is “the relation between an entity and its constitution” (Place, 1999, 81). It seems clear that a description of an object (e.g. a molecule of hydrogen oxide) and a description of its constituent atoms do not have a common referent. For one thing, there is only one molecule being described but there are three atoms being described.

Place would likely reply that he's not claiming that an object is identical to an unassembled collection of its constituent parts.

> Provided we specify their form and arrangement we can equally well say that the parts of a thing so arranged are the same thing as the thing itself and that the thing itself is the same thing as the collection of its parts arranged in that particular way. (Place and Schneider, 2013)

Here, Place says something quite sensible about (material) constitution; namely, that a material object is a combination of matter and form.

One might say that a molecule of hydrogen oxide is constituted by two atoms of hydrogen and one atom of oxygen; but, those atoms have to be arranged – chemically bonded together in a specific way – before they will constitute that molecule. If those three atoms are just floating around in an otherwise empty jar

---

25 Evnine (2011) claims that "Constitution is the relation between something and what it is made of. Composition is the relation between something and its parts." He considers various theories as to their relation and concludes that they are not related.

without being bonded together, they would not constitute a molecule of hydrogen oxide.

The molecular form is not just an arbitrary fact. The three atoms that, when bonded together, constitute a molecule of hydrogen oxide have to give up some energy to achieve that form; so, achieving it is an exothermic event in the lives of the atoms that participate in an occurrence of such an event.

From this perspective, it seems reasonable to say that, when two hydrogen atoms and one oxygen atom combine to form (constitute) a hydrogen oxide molecule, the object that comes into existence is identical to the atoms that make it up while they continue to exist in the form they must assume to constitute a molecule of hydrogen oxide. That object, that particular molecule of hydrogen oxide, is self-identical whether described as one molecule or as three particular atoms bonded together in such a way as to *be* that molecule of hydrogen oxide.

#### §3.3.3.4.4.1 Is an Afterimage a Material Object?

However reasonable this perspective may be as a theory of material objects (and I think it has considerable merit), it poses a challenge for Place's claim that the brain/consciousness relation is the relation between an entity and its constitution; particularly, when consciousness is understood as experience. It is not clear that it makes any sense to say that an experiential phenomenon such as the color of an afterimage is a material object composed of elementary particles of matter arranged in some form however complex.

Presumably, a microphysical description of physical events in the brain would include an account of neural firing patterns, the flow of calcium ions in the brain and other physical phenomena. While it's quite sensible to say that the neurons that compose a neural firing pattern are themselves composed of atoms and molecules arranged in a certain way, it is hard to see how such things as atoms, molecules and neurons could be arranged so as to constitute – to *be* – something other than another material object. So, unless the color of an afterimage is itself a material object, it wouldn't be composed of elementary particles and/or objects composed of elementary particles arranged in some form however complex.

To put it another way, an arrangement, no matter how complex, of elementary particles and/or material objects constituted by elementary particles will always be something more than just an appearance to an experiencing I, which is all that the color of an afterimage is.

#### §3.3.3.4.4.2 The Criterion of Success

Another problem with the constitutional view is that the identity of a thing and its constituents appears to be a token identity, a claim about particular things, particular molecules and their constituents. Place makes a strong effort to discredit token identity theories. Following Boring (1930), the originator of the identity theory, Place defines a “perfect” correlation as an identity.

> If, as seems more than likely, future research using the recently discovered techniques of brain imaging will allow us to identify such perfect correlations between mentally and

> physically specified variables, we shall be in a position to assert with confidence that at least some specifiable type-identity statements involving mentally and physically characterized processes are known to be true. In that case, who will give a fig for token identity physicalism? (Place, 1999, 89)

Here, I think, we have the weakness of type-identity theories: they rely on dubious criterion of success. If A invariably causes B and nothing else ever causes B, there will be a perfect correlation between A and B; but, there will be no identity. If A causes B, A has a property B lacks, being able to cause B. So, they couldn't be identical.

Even if scientists establish a perfect correlation between a physical phenomenon and an experiential phenomenon, if they fail to falsify identity theory or all non-identity theories, philosophers will be left to debate that issue. But, in that case, it would be difficult to say that identity theory is any more (or less) “scientific” than non-identity theory.

And still unaddressed is the logical problem facing a theory of brain/experience identity: something in the brain is held to be nothing more than an appearance to the subject of the experience. At least some attempt should be made to show that an experiential phenomenon such as the color of an afterimage is itself something more than an appearance to its subject.

My description of my experience of seeing an afterimage in terms of phenomena would mention an instance of greenness and squareness that appeared to float in the space between my eyes and the wall. The microphysical description would presumably mention neural firing patterns and/or other goings on in my brain.

*What is the one entity that these descriptions are both descriptions of?*

The crux of Place's theory is that we have two descriptions and we subsequently learn that they have a common referent. If that is the claim, I want to know what that common referent allegedly is.

Instead of responding to this question, Place might argue that I'm unfairly characterizing his position; and, he could be right. It's not at all clear to me that Place would allow one of the two descriptions alleged to be descriptions of a common referent to be a description of experience in phenomenal language.

> If we assume, for example, that when a subject reports a green after-image he is asserting the occurrence inside himself of an object which is literally green, it is clear that we have on our hands an entity for which there is no place in the world of physics. In the case of the green after-image there is no green object in the subject's environment corresponding to the description that he gives. Nor is there anything green in his brain; certainly there is nothing which could have emerged when he reported the appearance of the green after-image. Brain processes are not the sort of things to which colour concepts can be properly applied. (Place, 1956, 51)

Now, the phrase “the occurrence inside himself of an *object* which is literally green” (my emphasis) suggests that Place denies that an afterimage is a classical, Russellian sense datum. If so, I would agree. There aren't any of those.

Furthermore, I also agree with Place's other three claims: that there need be no green object in the subject's environment; that nothing in the brain is green; and, that color terms are not properly applied to brain processes.

Nevertheless, there is an instance of greenness that occurs to me; and, it must be accounted for. In my view, it exists in my experience, instantiated by its occurrence. It is one aspect of my experience singled out by the attention I put on it from the seamless web of ongoing experience that Searle calls the unified conscious field.

Place rejects with all of this. He denounces the concept of the “phenomenal field” (which I take to refer to ongoing phenomenality), calling it “mythological”. However, if Place denies the existence of the experiential phenomena that make up our ongoing experience – whether we call ongoing experience a unified conscious field or a phenomenal field or something else – he has adopted eliminative materialism packaged up as identity theory.

### §3.3.3.4.5 A Comment on Competing Identifications

Resolution of the controversies swirling around the Hard Problem of experience, knowledge arguments and the like is being impeded by an inability to recognize that the question is one of competing identifications, the materialistic and the phenomenalistic understanding of the referent of "experience".

The materialist, including the so-called identity theorist, identifies experience as (assumes that "experience" refers to) some sort of brain activity; and, assumes that the experiential phenomena associated with that brain activity either doesn't exist at all or may be safely ignored by the philosopher of consciousness.

The phenomenalistic interpretation assumes that experience exists and that it is constituted by the experiential phenomena that occur to (are instantiated in the experience of) the subject of the experience.

### §3.3.3.5 Is a Relevant Identity Theory Plausible?

Having shown that many so-called identity theories are not relevant to KA:TNG in virtue of their identity claims, he question that naturally arises at this point is: *What sort of identity theory is relevant to KA:TNG in virtue of its identity claims?*

Relevant, in my view, are claims of physical/phenomenal identities. I, of course, speak about phenomena when claiming that physical and experiential phenomena are not identical. In contrast, many philosophers speak about properties; for example, David Papineau writes:

> ... there is something very counter-intuitive about the phenomenal-material identity claims advocated by materialists. When materialists urge that *seeing red* (and here you must imagine the redness) is identical to some material *brain property*, it strikes many people that this *must* be wrong. From now on I shall call this natural reaction the 'intuition of mind-brain distinctness'. (Papineau, 2002, 74)

*How does Team Identity aim to make its point?*

Clearly, Papineau and I are in general agreement as to the sort of claim that is relevant, a claim of physical/phenomenal identity; but, he's playing for team Identity and I'm playing for team non-Identity. However, because he speaks of properties whereas I speak of phenomena, Papineau and I would likely each accuse other of making a category error, mistaking a phenomenon for a property

or vice versa.

We're not only playing for opposing teams, we're playing by different rule books, so to speak. The practical effect of this choice is that team non-Identity has the appearance/reality distinction available to help defeat an identity claim. Identity theorists who insist on speaking in the language of properties would need to show that phenomenal redness, say, is itself a property rather than a phenomenon; otherwise, they are not actually defending themselves against the argument for non-identity that phenomenon dualists are making.

*How plausible is the claim that experiential phenomena are identical to physical phenomena?*

It has been observed that “contemporary materialists eliminate mental individuals and reduce only states and events” (Lycan, 1987, 17). William G Lycan revisits the venerable argument from illusion to explain the problem and to show that representationalism was designed to solve it.

> Sensory qualities pose a serious problem for materialist theories of the mind. For where, ontologically speaking, are they located? Suppose Bertie is experiencing a green after-image as a result of seeing a red flash bulb go off; the greenness of the after-image is the quale. (Lycan, 2015, §1)

Except for the cause of the afterimage, Bertie's situation presents the same issues as is presented by Mary and her *otamot green* afterimage. The problem arises from the intuition, *I experience; therefore, I experience something,* a modernized version of the intuition attributed to Plato: “He who sees must see something.” (Anscombe, 1965, 68)

> … a materialist might suggest a type-identity of Bertie’s phenomenal greenness with something neurophysiological, but it is not plausible to think that a smoothly and monadically green patch in one’s visual field *just is* a neural state or event in one’s brain. At best, the type-identity theorist would have to do away with the important claim that greenness itself, rather than some surrogate property, figures in Bertie’s experience; the suggestion would be an error theory, and would have to explain away the intuition that, whatever the ultimate ontology, Bertie really is experiencing an instance of greenness. (Lycan, 2015, §3.2)

Lycan is certainly correct to say that type-identity theorists do away with the greenness itself, considered as an experiential phenomenon. Lewis and Harman ignore intentional objects; Byrne and Hilbert deny their existence.

The existence of an instance of greenness poses a problem for materialism, a problem defined by the following argument which Lycan considers to be valid (as do I).

> [LAM-1] Bertie is experiencing a green thing.
>
> [LAM-2] Suppose that there is no physical green thing outside Bertie’s head. But
>
> [LAM-3] There is no physical green thing inside Bertie’s head either.
>
> [LAM-4] If it is physical, the green thing is either outside Bertie’s head or inside it. Thus,

[LAM-5] The green thing is not physical. [1,2,3,4] Thus,

[LAM-6] Bertie’s experience contains a nonphysical thing. [1,5] Thus,

[LAM-7] Bertie’s experience is not, or not entirely, physical. [6]

Disputes about the soundness of Lycan's Argument against Materialism, LAM, turn on the quiddity – the whatness – of that which is experienced.

Lycan notes that “the materialist cannot admit that the greenness of the after-image is a property of an actual sense-datum” because Russellian sense data are supposed to be mental objects with whose (presumably mental) properties the experiencer becomes acquainted, thereby accounting for the experience of greenness. If there are any such things as sense data in the Russellian sense, materialism is false.

Lycan sums up the difference between the representationalist approach and the sense data theory, thus: “In defending his sense-data, Russell mistook a nonactual material thing for an actual immaterial thing”. Thus the green thing of [LAM-1] and [LAM-5] is a non-actual thing. Invoking the usual possible worlds semantics for these non-actual objects, Lycan locates them as actual objects in possible but non-actual worlds.

Such a theory may provide materialists with a defense to the argument from illusion; but, I find explaining experiential phenomena as nonactual material objects just as unpalatable as explaining experiential phenomena as actual mental objects. In my view, an instance of phenomenal greenness such as Bertie experiences is an actual experiential phenomenon occurring to the actual Bertie living in the actual world rather than to a counterpart of Bertie in some non-actual world.

#### §3.3.3.5.1 Is This a Theory of Sense Data?

I see no reason why the concept of a sense datum can't be weakened so that an experiential *phenomenon* can be a sense datum despite not being either a mental *property* or a mental *object* whose properties somehow account for our experience; and, I'm okay with considering phenomenon dualism to be a theory of sense data in this weakened sense.

According to Lycan,

> … phenomenal individuals such as sense-data are intentional inexistents a la Brentano and Meinong. It is, after all, no surprise to be told that mental states have intentional objects that may not exist. So why should we not suppose that after-images and other sense-data are intentional objects that do not exist? If they do not exist, then – *voila!* – they do not exist; there are in reality no such things. (Lycan, 1987, 88)

Now, a phenomenon dualist would certainly agree that there are no afterimages in *physical* reality. The disputed question is whether phenomenal reality is distinct from physical reality.

We have come full circle. We are back to the dilemma Dennett faces in trying to classify an afterimage as an intentional object that doesn't exist.

If the non-existence claim is that the instance of phenomenal greenness that

constitutes Bertie's afterimage is an intentional object that does not exist as a physical object, I agree. However, if the non-existence claim is that the afterimage is an intentional object that does not exist in any sense whatsoever, I disagree. An afterimage can't be an intentional object without being something rather than nothing at all.

Some clarification of the term “intentional inexistence” is sorely needed. Lycan is apparently interpreting that term as a synonym for “non-existence”. I would allow intentional inexistence as a mode of existence rather than as a mode of non-existence. In my view, an experiential phenomenon such as Bertie's afterimage has intentional inexistence because it is instantiated in or exists in Bertie's experience despite not being instantiated as a physical phenomenon.

*How is this version of sense-data any better than the original?*

This version lowers the stakes of the philosophical debate.

According to Lycan, “If there (really) are phenomenal individuals such as sense-data, then materialism is false right there”. (Lycan, 1987, 18) If sense data of the Russellian sort exist, Lycan's unqualified statement is true. No form of materialism or physicalism would survive.

However, my claim is that experiential phenomena such as afterimages are phenomenal individuals despite not being sense data in the Russellian sense. As long as the distinction between experiential *phenomena* and mental *objects* is kept in mind, an experiential phenomenon may be considered a sense datum in a weaker sense than is attributed to Russell.

In that case, though, only those forms of materialism and physicalism that constitute phenomenon monism would conflict with phenomenon dualism; but, they would be in conflict whether or not phenomenon dualism is considered a theory of sense data.

### §3.3.3.4.2 Is Representationalism Unmotivated?

Lycan argues that representationalism was intended to defend materialism from the sense data theories that seemed to follow from the argument from illusion; but, surely, that claim requires qualification. Some representationalist theories (e.g. Jackson, 1977) are quite friendly toward sense data in the Russellian sense.

Certainly, some representationalist theories defend materialism by eliminating sense data in the Russellian sense; but, it is not clear whether they must also be eliminative as to experiential phenomena which are sense data only in a weakened sense of that term.

Clearly, phenomenon dualism is a challenge to forms of materialism that assume phenomenon monism; so, eliminative materialists and identity theory physicalists may still want to employ suitable forms of representationalism to defend their views; but, there is no logical reason why materialists and physicalists who are not phenomenon monists can't use other forms of representationalism to defend their views.

The point is that representationalism does not decide the question how many

fundamentally distinct kinds of phenomena there are. Advocates of different answers to that question are each free to adopt a compatible form of representationalism.[26]

### §3.3.3.4.3 Feigl, An Existential Identity Theorist

Herbert Feigl was one of the founders of identity theory; but, his views are very different from those of Place and Smart on at least one key point: Feigl expected an identity theory to identify the referent of terms in the phenomenal language used to describe experience with terms in the neurophysiological language used to describe brain activity.

The terms in the phenomenal language designate by directly labeling the qualia or "raw feels" of immediate experience, qualities within the phenomenal field of which we are immediately acquainted. The terms in the neurophysiological language designate "complicated, highly ramified patterns of neuron discharges" (Feigl, 1958, 5e).

Although scientists will continue to provide further evidence upon which to base the identification, it ultimately requires a choice on the part of the philosopher.

> We have stressed that the (empirical!) identification of the mental with the physical consists in regarding what is labeled in knowledge by acquaintance as a quale of direct experience as identical with the denotatum of some neurophysiological concept. The scientific evidence for parallelism or isomorphism is then interpreted as the empirical basis for the identification. The step from parallelism to the identity view is essentially a matter of philosophical interpretation. (Feigl, 1958, 5e)[27]

Feigl notes the distinction of evidence and evidenced. As he uses it, this distinction is equivalent to the appearance/reality distinction. The appearance of a physical object such as a brain is the evidential basis for our judgment that there is a physical object beyond the appearance; consequently, they are not one and the same.

Feigl tells us that

> If the denotatum of "brain process (of a specified sort)" is confused with the appearance of the gray mass of the brain as one perceives it when looking into an opened skull, then it is indeed logically impossible to identify this appearance with the raw feels, e.g., of greenness or of anxiety. (Feigl, 1958, 5e)

This is something of a straw man argument, though. How many philosophers are attempting to identify the appearance of the brain with an instance of greenness? The existential identity theorists is attempting to identify some neural activity – not the appearance of that activity – as being nothing other than an instance of greenness. That instance of greenness also is an appearance, to be sure; but, it's a different appearance; and, it still seems logically impossible that

---

26 Torin Alter makes a similar point in his critique of Jackson's attempt to undermine the KA. “In the debate over the knowledge argument, representationalism would appear to be a red herring.” (2007, 74)

27 If it comes down to a matter of choice, we would also have the choice of whether to see this outcome as a vindication of mysterianism rather than of physicalism.

it be identical to something – some pattern of neural activity – that is not an appearance at all.

### §3.3.3.6 How to Justify a Relevant Identity Claim

In principle, identity theorists could attempt to establish their identity claims by applying Leibniz's Law of Identity of Indiscernibles to evidence showing that the phenomenal kind and the material kind have exactly the same set of properties. In §3.3.3.6.1 I will argue that such an effort would fail.

In practice, identity theorists attempt to establish their identity claims in a variety of other ways, including: (1) as an inference from isomorphic sets of facts; (2) by way of a consideration of causal functions; and, (3) by a claim about a posteriori identities.

In §3.3.3.6.2, I will consider Austen Clark's identity theory as way to assess the relevance of theories that infer an identity from isomorphic sets of facts.

In §3.3.3.6.3, I will consider the Phenomenal Concepts Strategy for presupposing an identity claim.

In § 3.3.3.6.4, I will consider the claim that scientists will someday reveal that physical/phenomenal identities are a posteriori necessities the way they've already (allegedly) shown that water = $H_2O$ is a necessary a posteriori truth.

In §3.3.3.6.5, I will assess the relevance of identity theories that focus on causal functions, focusing on Papineau's causal argument for materialism and, in §3.3.3.6.6, Jackson's “Mark I” type-identity theory.

#### §3.3.3.6.1 The Appeal to Leibniz

In preparing the ground for an evaluation of identity claims, we must first clarify the criteria by which to decide whether "... whether two independent sets of observations are observations of two different sets of correlated events or of one and the same process or event." (Place and Schneider, 2013)

The two sets of observations are, of course, third-person observations of physical phenomena in the brain and first-person experiences of experiential phenomena. In short, third-person phenomenology and first-person phenomenology.

Place and Schneider explain how an identity claim might be established. In principle, advocates of a claim that some item, A, is the very same item as B could appeal to Leibniz.

> Leibniz's principle of the Identity of Indiscernibles holds that if all the predicates that are true of an entity A are also true of what is taken to be another entity B and all the predicates that are true of B are also true of A, then A and B are not two things but one and the same thing. (Place and Schneider, 2013)

Establishing an identity in this way is a hopeless task because ...

> … there are a number of properties which apply to all brain processes which the introspecting subject would never think of predicating of his experiences solely on the basis of his having or experiencing them. One such property is the property of involving the firing of at least one and probably many thousands of neurons each of which has a

> specific location within the anatomical structure of the brain. (Place and Schneider, 2013)

To a non-identity theorist, this passage concedes that brain processes and experiences have different sets of applicable predicates; consequently, by an application of Leibniz's Law of Identity (the Indiscernibility of Identicals), it is the basis for a conclusion of non-identity.

Leibniz's Law “... holds that if two descriptions A and B refer to one and the same entity, then any predicate which forms a true proposition when predicated of A must also form a true proposition when predicated of B” (Place and Schneider, 2013).

The strategy that opponents of identity claims use is simple: show that some predicates attributable to A are not attributable to B or vice versa. Then, by modus tollens, one may conclude that A and B are not one and the same entity.

For one example, William S Robinson argues for the distinctness of physical and experiential phenomena on the basis of their having different sets of properties.

> Qualia realism holds that certain neural activations cause our experiences to have qualia such as green, sweet, peppery, cold, pain, itchiness, nausea or sexual pleasure. On the one hand, no one would suggest that these properties are literally had by sets of neural activation events. And, on the other hand, nothing appears to us in ordinary experience as having the kind of complexity that neural events actually have, according to our best neuroscience. Thus, qualia realists easily distinguish between qualia (= qualitative properties of our experiences) and properties of neural activations. (Robinson, William S. 2013b)

While we don't have an impasse, we do have a situation in which identity theorists, unable to establish their claim on logical grounds, focus on defending themselves from arguments for non-identity. To this end, Place and Schneider consider three arguments against identity theory. Each of these arguments appeals to Leibniz to show that an (apparent) difference in properties justifies a conclusion of non-identity. Identity theorists defending against these arguments for non-identity aim to show that the conclusion of non-identity does not follow despite the (apparent) lack of a common set of properties.

#### §3.3.3.6.1.1 The Argument from Phenomenal Properties

> The argument from the phenomenal properties of experience may be stated as follows: experiences have phenomenal properties such as the property of being green, red, blue or yellow. It makes no sense to describe a brain process as green, red, blue or yellow. Hence experiences have properties (phenomenal properties like being of a certain colour) which no brain process can have. Hence by Leibniz's Law experiences cannot be the same thing as brain processes. (Place and Schneider, 2013)

This argument is easily recognized as a form of the argument that motivated Dennett to deny the existence of the red stripe in his experience after he induced a flag afterimage.

*How does the identity theorist defend against the threatened conclusion of non-identity?*

### § 3.3.3.6.1.2 The Topic Neutrality Defense

Place initially says that the argument from phenomenal properties breaks down because it ignores what Smart (1959) called the topic neutrality of our descriptions of our private experiences.

I readily agree that, whenever possible, the language of discourse should not be allowed to prejudice outcome of the inquiry; but, Smart's handling of topic neutrality was flawed *ab initio*, from the getgo.

Here is Smart's suggestion

> My suggestion is as follows. When a person says, "I see a yellowish-orange after-image," he is saying something like this: "*There is something going on which is like what is going on when* I have my eyes open, am awake, and there is an orange illuminated in good light in front of me, that is, when I really see an orange." (Smart, 1959, 149)

The problem here is that "There is something" is singular in number; and, far from being topic neutral, assuming at the outset that there is only one something involved is question begging. It prejudices the inquiry into how many somethings (events, phenomena, processes or whatever) need to be described to give a full description of all that is happening.

I don't mind stating the topic under discussion in a way that is outcome neutral; but, in my view that requires (at least) two statements: a statement about what is going on in my experience and a statement about what is going on in my brain.

Just having two statements does not entail either that they are or that they are not descriptions of the same phenomenon or the same event or process or whatever. So, when I am experiencing an greenish afterimage, I insist on saying *both*

1. That here is something going on in my experience that is like what goes on in my experience when I have my eyes open while awake and there is an unripe tomato illuminated in good light in front of me; and,
2. That there is something going on in my brain that is like what goes on in my brain when I have my eyes open while awake and there is an unripe tomato illuminated in good light in front of me.

Now, if identity theorists wish to present their evidence and/or their arguments for the claim that what is going on in my experience is identical to what is going on in my brain, they are free to do so.

My conclusion that there are two phenomena or two events to be described is based on the argument that the instance of phenomenal greenness that I experience is an experiential phenomenon, an appearance to me; whereas, the instance of electrical activity going on in my brain is an instance of a physical phenomenon. The one is a phenomenal reality to me, its experiencer; and, it has a subjective or first person mode of existence. The other is a physical reality; and, may be measured or detected by anyone with suitable training and equipment.

They can't possibly be identical.

### § 3.3.3.6.1.3 The Color Realism Defense

> … when we describe an experience such as an after image as green, we are not predicating the property of greenness to the experience itself, we are saying only that the experience is the sort of experience we normally have when we look at objects or perceptible phenomena which do have the property of being green. In other words visual experiences do not have the property of being literally green any more than brain processes have this property. Leibniz's Law is not infringed. (Place and Schneider, 2013)

I can agree that we are not predicating greenness of experience; but, my reasons are different. I do not want to make experience a property bearer; so, I deny that greenness is a property of experience. However, I also deny the assumption of color realism that Place introduces by assuming that ordinary objects can literally *be* green or red or some other color.

> This argument presupposes of course, that colour words like 'red', 'green', 'blue' and 'yellow' when used literally, refer to physical properties of objects and phenomena in the external physical world and are not, as has been traditionally supposed, the names of certain properties of the individual's subjective visual experience. (Place and Schneider, 2013)

The obvious problem with this approach is that Place classifies his theory as a scientific hypothesis; yet, as color realists Byrne and Hilburt (2003) readily admit, color scientists mostly reject color realism.

The less obvious problem is that Place cites only linguistic evidence for color realism; for example, the fact that we describe objects like leaves as green and tomatoes as red. Let us assume for the sake of the argument that most people would be willing to affirm "the tomato is red" when shown a ripe tomato and that this counts as an empirical fact.

How do we get from that fact to a conclusion about the actual (scientifically detectable) properties of the tomato?

If we assume that the is in "the tomato is red" is the is of predication, it follows that a sentence of this construction ascribes a predicate to the tomato. Now, as it happens, the name of a one-argument predicate is 'property'. To ascribe a predicate is to ascribe a property; so, to say "the tomato is red" is to say that the tomato has the property of being red or the property of redness.

This conclusion is contrary to the scientific perspective which holds that objects don't have color properties; so, we need to uncover the flaw(s) in this argument.

One flaw is the assumption that predicate ascription entails property detection, the automatic assumption that the world conforms to our manner of speaking about it. The antithesis of such a claim would be that constructions like "the tomato is red" employ a figure of speech.

Parents and teachers may well use constructions like "the tomato is red" when teaching children the proper use of color terms. We point to objects and say "that [tomato] is red", "the sky is blue", "those blocks are green" and so on.

To facilitate philosophical reflection on the brain/experience relation we need to outgrow whatever faulty ideas we acquired about color properties from the way

children are taught to use of color terms. Otherwise, as philosophers, we'll end up using a figure of speech to trump the scientific perspective; and, that is something I refuse to do.

There is also another, more subtle flaw, in the pseudo-argument for color realism: the assumption that we must use the is of predication.

When I say "the tomato is red", I'm using the is of appearance rather than the is of predication. By "the tomato is red" I mean "the tomato appears red to me".

### § 3.3.3.6.1.4 Is Greenness Phenomenal Greenness?

Place admits that experience can have the property of phenomenal greenness; but, counters that phenomenal greenness can be predicated of brain processes as well.

Here, I think, is the point at which to raise the possibility that the identity theorist's defense to the argument from phenomenal properties gets an illegitimate boost from the commonly made mistake of treating phenomenal greenness as a property rather than a phenomenon.

I deny color realism as antiscientific and as property dualistic. In particular, I reject the assumption on which Place relies, that color words like 'red' and 'green' refer to physical properties of objects and phenomena of the external physical world. Rather, in my view, such terms are the names of experiential color phenomena.

So, as I use the term, "green" is a synonym for what Place calls "phenomenal greenness".

Place seems to acknowledge this fact when discussing the history of the identity theory. He attributes the first statement of identity theory to Boring (1930)

> To the author a perfect correlation is identity. Two events that always occur together at the same time in the same place, without any temporal or spatial differentiation at all, are not two events but the same event. The mind-body correlations as formulated at present, do not admit of consideration as spatial correlation, so they reduce to matters of simple correlation in time. The need for identification is no less urgent in this case. (p.16).

However, in explaining why it took a quarter century before identity theory gained much traction, Place writes that Boring was

> ... apparently committed to combining the identity theory with a phenomenalist account of sensory qualities which on Leibniz's principle of the Identity of Indiscernibles would commit him to the view that certain brain events are literally green, high pitched, warm, sour or putrid, which for a philosopher would constitute an immediate knock-down reductio ad absurdum of his position. (Place and Schneider, 2013)

Now, I have a phenomenalistic view of sensory qualities; meaning, I hold that the qualities we experience – redness, sweetness and so on – are experiential *phenomena*. And, relative to this frame of reference, Boring's position is absurd, just as Place and Schneider say.

So, the argumentative burdens of identity theorists and non-identity theorists

can be stated as follows:

1. Identity Theorists must jettison the understanding of sensory qualities as phenomena.
2. Non-Identity Theorists need only maintain a phenomenalistic view of sensory qualities.

### § 3.3.3.6.1.5 The Privacy of Experience

Place argues that our experiences are private to the experiencer by the way we know about them. Since knowing is an intentional verb, one can't argue that experiential phenomena are private, physical phenomena are not private; therefore, they can't be identical.

That knowing is an intentional verb is uncontroversial. Place and Schneider give a good example of why.

> If Joe knows that James has red hair and James is the brother of John, it does not follow that Joe knows that John's brother has red hair. This follows only if Joe also knows that James is the brother of John.

However, it's not at all obvious that "experience" is also an intentional verb when used in a statement such as "I am experiencing an afterimage". If I am right, an appearance is only as it appears; so, experiencing any experiential phenomenon is experiencing an appearance and not something I don't experience as such.

Nevertheless, for the sake of the argument, let us suppose that "experience" is an intentional verb. The identity theorist's argument then seems to be that the following argument is fallacious due to what Churchland called the intensional fallacy.

[FIF-1] I am experiencing an afterimage

[FIF-2] I am not experiencing any brain activity

[FIF-3] An afterimage is not identical to any brain activity

The identity theorist replies as Churcland replied to Nagel and Jackson:

> The reason for this perhaps, is that what is at issue in the case of privacy is not what a man knows about his experience as compared with what he knows about his brain processes, but how he comes to know what he knows in the two cases. (Place and Schneider)

Place's defense to the argument from privacy presents us with the same situation as does the NCP (Nagel/Churchland/Polanik) argument considered in §3.3.3.3.3. It is uncontroversial that there are two ways of knowing involved. The question is whether what is known via one way of knowing is identical to what is known via the other way of knowing.

Now the fallacy of the Fallacious Intensional Fallacy, FIF, argument stated above is that the identity theorist appears to expect the non-identity theorist to rely solely on what is known by examining or reflecting on his or her own subjective experience to reach a conclusion as to whether what is known by one way of knowing is identical to what is known by the other way of knowing.

I accept no such limitation.

Both [FIF-1] and [FIF-2] are based on what I know by examining and reflecting on my own subjective experience. I could, with the utmost sincerity, issue both statements from the comfort of my armchair without examining my brain or letting it be examined by anyone else. But, once it is admitted that there is another way of knowing involved, knowing by scientific description, I insist on making use of it as well.

Now, *how* I know something via some way of knowing certainly tells me *something* about *what* is known via that way of knowing; otherwise, there is no reason to call it a way of knowing.

Among the facts that we know from scientific investigations into the nature of physical phenomena occurring in the brain is that physical phenomena are physical realities; meaning, that they exist objectively and are objectively measurable/detectable with scientific instruments. However, once we understand an experiential phenomenon as an appearance to a subject, we have a problem.

Identity is a symmetric relation. If A and B are the same item, then A is nothing other than B and B is nothing other than A. If we are unwilling or unable to affirm both such claims, we don't have an identity relation.

Would we say both that an objectively physical phenomenon occurring in the brain is nothing other than an appearance to a subject; and, that an experiential phenomenon – an appearance to an experiencing subject – is nothing other than an objectively detectable physical phenomenon occurring in the brain? I don't think so.

An eliminative materialist might try to deny the existence of any and all appearances. This is Dennett's strategy with respect to afterimages and, presumably, all other experiential phenomena. But, if that path is not taken, the identity option becomes implausible; given that an afterimage is a phenomenon, who would say that a physical phenomenon occurring in the brain while an afterimage is experienced is nothing other than an appearance to a subject?

### § 3.3.3.6.1.6 The Absence of Spatial Location

Place seems to be saying that an experiential phenomenon such as an afterimage is a process; and, that every process involves "continuous change which is extended over time" and "cannot be said to exist or to be going on unless it is going on somewhere".

I have no objection to the first criterion. An afterimage fades away over time; and, it seems reasonable to say that such a change is continuous.

Problematic is the second criterion. It is question-begging as to the nature of an experiential process.

Certainly, while a physical phenomenon is occurring, it must be occurring somewhere. But, while an experiential phenomenon is occurring it must be occurring to some experiencing I. For a physical phenomenon such as a neural firing pattern, occurrence is occurrence at a place; so, a physical phenomenon

that is occurring exists at a location while it is occurring at that location.

With experiential phenomenon, occurrence to an experiencing I locates it in the experience of that experiencer; and, we could reasonably say that an experiential phenomenon exists while it is occurring in the experience of some experiencing I; but, it is not at all clear that "in the experience of an experiencing I" is anything other than a metaphorical location.

Physical phenomena certainly occur at a place; they are measurable or detectable at that place. Certainly, one could *assume* that some particular physical phenomenon occurring in the brain is identical to an appearance to an experiencing I; and, if we did that, it would surely follow that the afterimage has a physical location within the physical universe.

However, that assumption is simply question begging in a debate over *whether* some physical phenomenon is identical to an experiential phenomenon. Without the assumption of identity, we have no grounds for concluding that an afterimage has a physical location.

Suppose that we all adopted the convention of assigning  experiential phenomena the location of the physical phenomenon with which it is empirically correlated. We could say that an afterimage is located at the location in its experiencer's brain where a correlated neural firing pattern is located.

Such a convention, even if reasonable, would not change the fact that no such convention is required for physical phenomena to have an actual, non-metaphorical physical location. So, the intuition that physical and experiential phenomena are distinct kinds of phenomena remains intact.

#### § 3.3.3.6.1.7  Evaluating the Defenses

Identity theories are often defended as being simpler than dualistic alternatives; but, it is clear that identity theories are not so simple if one considers the claims that are tacked onto them in defending against arguments for non-identity; for examples, the claims that

1. The sensory qualities commonly known as qualia are properties rather than phenomena;
2. Color realism (and, presumably, quality realism in other sense modalities) is true; and,
3. The insistence that the issue is exclusively *how* something is known rather than *what* is known in experiencing.

When these issues are squarely faced, non-identity theory may turn out to be the simpler theory.

### §3.3.3.6.2  Identity via Parallel Phenomenology

Commenting on arguments that purport to show that the physical does not entail the phenomenal, Austen Clark wrote

> What the various "conceivability" arguments show is not that the identities are false, but rather they cannot be explained. They retain an arbitrary character. We cannot explain why process G is associated with one particular qualitative content instead of some other one. (Clark, 1994)

Clark then argues that trying to demonstrate an identity between, say, green and some brain process, G, is a misguided effort. Instead he proposes to discover the relations of similarity and differences obtaining between experiential phenomena. For instance that red is more similar to orange than to yellow; or, that some particular shade of green is the complement to some particular shade of red.

"It is facts of this sort that will link the physical story P to the appearances presented to Joe", writes Clark. One would then map the resulting *quality space* to neurological activity

> But although there seem to be no individual necessities – no particular necessity in attaching this quale to that neural process – given the structure of our quality space, there is only one way that the two can fit together. If you had full knowledge of the opponent processes – of Joe's Y-B, R-G, and Wh-Bl systems – and of the phenomenal structure of color quality space, then you would see that there is only one possible mapping. (Clark, 1994)

The relation between any one experiential phenomenon such as *tomato red* and the corresponding physical phenomenon will still seem like an arbitrary identity; but, when embedded in a quality space and mapped to neural phenomena, the particular identities will seem more reasonable, says Clark.

> We can explain why particular identities seem so arbitrary – as if the two phenomena are just arbitrarily stuck together. If one examines a particular identity in local terms, it is arbitrary. There is nothing in the concept of red or in the neighborhood of red, or in squiggle-squiggle or neural processes similar to squiggle-squiggle, which serves to explain the identity of sensing redly with process squiggle-squiggle. It can only be justified in global terms. It is only the overall fit of the structure of similarities to neural mechanisms of discrimination which justifies the particular identities. (Clark, 1994)

It may well be that Clark's account makes the alleged identities between a particular experiential color phenomenon and its corresponding neural phenomenon seem less arbitrary; but, there are still a number of weaknesses in this approach.

### *The Appearance of Reality*

Clark admits that color phenomena are appearances presented to an experiencer. I agree with that; but, once one admits that the items under discussion are experiential *phenomena*, appearances to a subject, one has to wonder how a physical object such as the brain could be nothing other than an appearance to a subject.

### *The Nature of the Link*

According to Clark, the link between the scientific account and the appearances presented to the subject will be provided by facts about qualitative structure rather than by the more usual reductive conceptual analysis. He argues rather convincingly that our knowledge is or will be so detailed that there will be only

one way to map the quality structure to the physiological facts; and, I'm willing to assume arguendo that he is right.

But his argumentative burden is to provide “a derivation of the identity” claims; and, he hasn't done that. There is nothing in his argument that rules out the possibility that all he has done is to specify the locations in the nervous system where a causal process of some kind generates the experiential phenomenon associated with nervous system activity at that location.

Until the empirical evidence rules out all versions of identity theory or all versions of non-identity theory, the evidence merely pinpoints precisely which items the identity theorist must identify and the non-identity theorist must distinguish (by philosophical reflection and debate, presumably). But, in such circumstances, a philosopher's choice between identity theory and non-identity theory is not compelled by the empirical evidence.

The non-identity theorist may respond to Feigl's admission, discussed above, that “the step from parallelism to the identity view is essentially a matter of philosophical interpretation” by adopting a different interpretation. Debate would then involve comparing the attempts to explain this or that body of evidence.

Is there a plausible alternative to the assumption that physical/phenomenal correlations are due to physical/phenomenal identities?

#### §3.3.3.6.2.1 A Non-Identity Alternative

As a non-identity theorist, I find I am unable to ignore the similarities between Clark's idea of mapping quality space to neural activity and Chalmers' speculation that information has two aspects, physical and phenomenal.

> Physical realization is the most common way to think about information embedded in the world, but it is not the only way information can be found. We can also find information realized in our phenomenology. States of experience fall directly into information spaces in a natural way. There are natural patterns of similarity and difference between phenomenal states, and these patterns yield the difference structure of an information space. Thus we can see phenomenal states as realizing information states within those spaces.
>
> ...
>
> Any given simple color experience corresponds to a specific location within this space. A specific red experience is one phenomenally realized information state; a specific green experience is another.
>
> ...
>
> This treatment of information brings out a crucial link between the physical and the phenomenal: whenever we find an information space realized phenomenally, we find the same information space realized physically. And when an experience realizes an information state, the same information state is realized in the experience's physical substrate.
>
> Take a simple color experience, realizing an information state within a three-dimensional information space. We can find the same space realized in the brain

> processes underlying the experience; this is the three-dimensional space of neurally coded representations in the visual cortex. Elements of this three-dimensional space correspond directly to elements of the phenomenal information space. (Chalmers, 1996, 283-285)

To summarize Chalmers' proposal as a single principle, the Hypothesis of Phenomenally Instantiated Information:

> **[HPII] Some of the information that is physically instantiated in the brain becomes phenomenally instantiated in experience.**

Sometimes there is no phenomenality associated with neural information processing; but, sometimes there is. Here, I am concerned with the latter cases.

*Would a dual-aspect theory be dualistic?*

Even if there is no information loss or distortion as the information becomes phenomenally instantiated, the physical instantiation of information would not be identical to the phenomenal instantiation of the same information; so, in my view, a dual-aspect theory of information would be an instance of phenomenon dualism.

*How reasonable is it to infer that information can be instantiated phenomenally?*

Pain conveys to my attention the information that some part of my body is in distress. A headache will usually prompt me to take some action to relieve the pain. In most cases, I will select from among the pain relievers I have available and carry on with whatever I'm doing; but, when not at work, I will sometimes decide to take a nap instead.

In any case, as these situations are experienced, it appears to me that I notice the pain occurring to me and focusing my attention on it; and, that my response is not the automatic response of the “zombie within” (Place, 2000). Whatever information is physically instantiated in my brain is apparently not sufficient to determine my response. In these situations, or so it seems to me while I am in them, I am required to choose my response.

From these considerations, I conclude that my response is a choice guided by phenomenally instantiated information. Recognizing that what I am experiencing is pain rather than hunger prompts me to consider taking a pain reliever rather than eating some ice cream. Recognizing the location of the pain also helps guide my decision making process. A pain in my back or neck will often prompt me to see my chiropractor. A pain in my toe generally doesn't call for more than trying harder to remember the location of furniture while walking around in the dark.

### §3.3.3.6.2.2 Phenomenally Instantiated Information?

Before proceeding further, I would like to distinguish my Hypothesis of Phenomenally Instantiated Information, HPII, from the similarly sounding Hypothesis of Phenomenal Information that David Lewis discusses.

I agree with Lewis' concession that

> No amount of the physical information that black-and-white Mary gathers could help her know what it was like to see colors; no amount of the physical information that we might

> gather about bats could help us know what it's like to have their experiences; and likewise in other cases. (Lewis, 1988, 84).

Consequently, when Mary is first presented with, say, a ripe tomato upon her release, she acquires information about the way experiencing is for her in those circumstances, information which she may then report to philosophers and other interested bystanders.

The question that divides us is whether this information is physical information or phenomenal information.

*What is the difference between physical and phenomenal information?*

If we limit physical information to information that Mary can acquire from her lessons, it's possible to say that physical information is information about the physical phenomena, properties, objects and their interactions that her lessons are about. Clearly, Mary acquires other information from personally experiencing experiential phenomena; and, it seems reasonable to characterize information about experiential phenomena as phenomenal information.

However, this answer isn't available to Lewis because he is eliminative as to experiential phenomena, arguing that materialists believe that "the phenomenal aspect of the world" does not exist (Lewis, 1988, 95). Thus, Lewis' answer would have to be that there is no phenomenal information because there are no experiential phenomena to have information about.

In contrast, existential identity theorists could accept this way of distinguishing physical and phenomenal information. They could argue that, for any experiential phenomenon, there is a physical phenomenon to which it is identical; and, therefore, that information about experiential phenomena is information about physical phenomena. However, this is precisely the point at which identity theory physicalism is challenged by knowledge arguments. Mary learns something from her first encounter with technicolor reality. She acquires information about the way experiencing is for her.

In reply, it may be argued that Mary does not learn any new information from her first encounter with the tomato because there isn't any new information to be had. It's all old information, physical information she already knew about from her lessons. The so-called "new fact" is nothing more than some old information presented to Mary in a new guise when instantiated phenomenally.

*How is this claim a defense of identity theory physicalism?*

The new guise is itself the experiential phenomenon at issue. Mary's report about this phenomenon when she first experiences it is a new fact about the way experiencing is for her. It doesn't matter whether new or old information is instantiated in the new appearance, that appearance – the guise – is itself the experiential phenomenon that's new to Mary.

The new guise in which old information supposedly appears to Mary is just how that information appears to her when phenomenally instantiated in her experience.

According to HPII, the experiential phenomenon, tomato red, that Mary

experiences, is a phenomenal instantiation of (some or all of) the information that became physically instantiated in *her* brain for the first time when she looked at the tomato the first time. However, Mary already knew all about the physical instantiation of information from her lessons and her studies of the brains of other people while they were experiencing this or that experiential phenomenon. Arguably, then, Mary's situation is a case where previously known information is presented in a new guise.

Now, if the phenomenal instantiation of some information is not identical to the physical instantiation of that same information, the old fact in a new guise defense fails to defend identity theory physicalism from dualism. We would have phenomenon dualism rather than phenomenon monism.

It is ultimately up to scientists to falsify or confirm [HPII], the Hypothesis of Phenomenally Instantiated Information; but, it seems like a reasonable hypothesis to me; and, it may well turn out that information Mary already knew about from her lessons and from studying the brains of other people is presented to her in a new guise when she experiences phenomenal redness for herself.

What I deny is that any of this defends identity theory physicalism.

### §3.3.3.6.3 The Phenomenal Concepts Strategy

Scientists have found the neurobiological correlates of various kinds of experiences. For example, experiencing pain is correlated with undergoing C-fiber discharging. However, just knowing about correlations between an experiential phenomenon and a physical phenomenon doesn't explain how it happens that they are correlated.

Some philosophers assert the identity of experiential phenomena and physical phenomena; and, presumably, they have what they take to be adequate reasons for asserting that claim. Such an identity claim defends materialism from the possibility of dualism; but, this rhetorical function has a price: the identity becomes inexplicable, a brute fact.[28]

Ned Block (2003, 9) invites us to suppose "that cortico-thalamic oscillation (of a certain sort) is the neural basis of an experience with phenomenal quality Q". Now, the identity theorist can dissolve the Hard Problem by asserting the identity "Q = cortico-thalamic oscillation (of a certain sort)".

> I think there is something right about this answer but it is nonetheless unsatisfactory. What is right about it is that if Q = cortico-thalamic oscillation, that identity itself, like all genuine identities, is inexplicable. (Block, 2003, 9)

I agree with Block and others that an identity is inexplicable. If someone were to ask “Why is Venus Venus?” an adequate answer is simply “It just is”. However,

---

28 A further consequence of adopting brute fact identities is that some arguments against dualism are sacrificed. In Consciousness Explained, Dennett defends adopting “the apparently dogmatic rule that dualism is to be avoided at all costs. It is not that I think that I can give a knock-down proof that dualism, in all its forms, is false or incoherent, but that, given the way dualism wallows in mystery, accepting dualism is giving up.” (1991, 37) It is now clear that non-identity is what makes explanation possible, though not inevitable.

the problem with physical/phenomenal identity claims is not that they are inexplicable identities. The problem is that they are *unjustified* identities. Block noticed this as well.

What's *wrong* with answering the Hard Problem by asserting an identity is that

> The claim that Q is identical to cortico-thalamic oscillation is just as puzzling – maybe more puzzling – than the claim that the physical basis of Q is cortico-thalamic oscillation. We have no idea how it could be that one property could be identical both to Q and cortico-thalamic oscillation. (Block, 2003, 9)

I think Block is quite right here as well. To paraphrase McGinn (1989, 349), no one has the slightest clue as to how technicolor phenomenology could be identical to electrochemical activity in the brain.

*What's an identity theorist to do?*

Unquestionably, the identity theorist could present a case for the identity claim. I've already questioned the relevance of many identity claims; and, in §3.3.3.6.4, I will I consider attempts to justify a posteriori identities. At the moment, I want to discuss the alternative that Block pursues.

The identity theorist offers therapy for the perplexed.

> How could one property be both subjective and objective? Although no one can explain an identity, we can remove puzzlement by explaining how an identity can be true, most obviously, how it is that the two concepts involved can pick out the same thing. This is what we need in the case of subjective/objective identities such as the putative identity that Q = cortico-thalamic oscillation. (Block, 2003, 9-10)

Identity theorists defending against the dualistic conclusions of knowledge arguments by using the Phenomenal Concepts Strategy, PCS, generally make no effort to show that their identity claims are actually true. Their defensive strategy seems to be that of showing that they have a reply to the KA that deflects or blocks its dualistic conclusions. As Ned Block explains,

> In the room, Mary knew about the subjective experience of red via the objective concept *cortico-thalamic oscillation*. On leaving the room, she acquires a subjective concept *this [mental image] phenomenal property* of the same subjective experience. In learning what it is like to see red, she does not learn a new fact. She knew about that fact in the room under an objective concept and she learns a new concept of that very fact. One can acquire new knowledge about old facts by acquiring new concepts of those facts. (Block, 2003, 12-13)

To the extent that Block's statement is representative, PCS tacticians – my term for those who adopt the Phenomenal Concepts Strategy – simply *assume* that cortico-thalamic oscillation is identical to a subjective experience. Consequently, PCS rests on a conditional claim:

> **[PCS-CC] If physical phenomena are identical to experiential phenomena, then arguments to the contrary are unsound and advocates of those arguments are mistaken, possibly because they've been taken in by an illusion.**

At this point, PCS tacticians offer their speculations as to the origin of the illusion casting its spell over those who reject the identity theory; and, some of

those speculations may be interesting. However, without presenting some evidence or some argument to show that the antecedent of [PCS-CC] is actually true, PCS tacticians are resting their defense of physicalism on a conditional claim that may only be vacuously true.

#### §3.3.3.6.3.1 Reply to Block

I'm willing to concede that, in her room, Mary has a concept of a physical phenomenon, a third-person phenomenon that is objective because it is objectively measurable by anyone with the right equipment. I've been calling it NFP-4738 intending to refer to whatever turns out to be the relevant neural phenomenon. Maybe it will turn out to be cortico-thalamic oscillation; or, maybe, it'll be something else.

Upon her release, Mary acquires a concept of an experiential phenomenon, tomato red, a first person phenomenon that is subjective because it occurs to a subject, in this case, Mary. Although a team of researchers could each measure Mary's brain and detect the presence of NFP-4738, only Mary experiences the redness.

At some point, Mary forms concepts of the objective phenomenon and other concepts of the subjective phenomenon under discussion; but, it need not be the case that concepts of objective phenomena are very different from concepts of subjective phenomena. Indeed, there might be no difference at all. Even before her release, Mary could very easily have become proficient in the use of *terms* for experiential phenomena, including *tomato red*. When she experiences that phenomenon for the first time, she becomes acquainted with its referent, tomato red. The term itself, *tomato red*, may be as mundane as the term for tomato.

There are those who deny the existence of phenomenal concepts as PCS tacticians use that term (e.g. Ball 2008, 2009). Tye's (2009) denial may be of particular significance because he previously defended the existence of phenomenal concepts. Nevertheless, we have terms by which to refer to experiential phenomena; and, it may turn out that the concepts by which we refer to experiential phenomena are peculiar in various ways.

Some such claims are not controversial. For example, PCS tacticians and non-identity theorists can both support what Stoljar (2005, 471) calls the Experience Thesis:

> **[Experience Thesis] Subject S possesses the (phenomenal) concept C of experience E only if S has actually had experience E.**

This is a mild thesis that doesn't commit anyone holding it to the belief that a phenomenal concept is peculiar in some way that affects the debate over the identity thesis. A phenomenal concept may simply be a concept that refers to experiential phenomena.

From a phenomenological standpoint, we encounter what at first glance appear to be two kinds of phenomena. A first-person phenomenon is only experienceable by its subject. A third-person phenomenon is objectively measurable by anyone. In experience there is technicolor phenomenology in multiple sensory modalities;

whereas, in the brain there is only electrochemical activity.

### §3.3.3.6.3.2 Reply to Loar

> I will assume that the antiphysicalists' phenomenological and internalist intuitions are correct. The idea is to engage them over the central point, that is, whether those aspects of the mental that we both count as phenomenologically compelling raise substantive difficulties for the thesis that phenomenal qualities (thus understood) are physical properties of the brain that lie within the scope of current science. (Loar, 1997, 597)

Loar certainly alludes to the rhetorical tactic necessary to avoid perpetrating a straw man argument: taking the antiphysicalists' point of view as they present it before trying to show that no antiphysicalistic conclusion follows. However, in my view, taking the phenomenological intuition seriously requires acknowledging

1. That experience falls under the concept of *phenomenon* rather than the concept of *property;* and,
2. That we may refer to experiential phenomena (the items of which first person phenomenology consists).

With respect to the first point, Loar may be forgiven for casting his presentation in the language of properties. He's responding to opponents who also speak in that language. However, in my view, his arguments fail more obviously when the debate is conducted in the language of phenomena; consequently, to the extent that I am successful in showing this to be the case, opponents of the identity theory will have good reason for shifting the language of discourse to one based on phenomena.

Loar argues that physicalists should accept that Mary learns a new fact about color experiences upon her release from monochromatic confinement; but, only on an "opaque" reading of that claim. In saying that experiencing color phenomena is opaque, Loar seems to mean that Mary's acquaintance with an experiential phenomenon such as tomato red does not reveal anything about any brain activity that is correlated with an occurrence of tomato red. Experiential phenomena are, therefore, opaque – they lack transparency – as to their relation to physical phenomena occurring at the same time.

So far, there is nothing to which a non-identity theorist must object. Everyone can agree that experiential phenomena generally don't reveal much, if any, information about the relation obtaining between an experiential phenomenon and any associated physical phenomena.

In the case of Mary and the tomato, the experiential phenomenon, *tomato red*, does not reveal the nature of its relation to any associated physical phenomena. In particular, it does not reveal which neurophysical phenomena it is related to nor the nature of that relation.

Identity and non-identity theorists part company concerning the thesis of revelation, [TR], that an experiential phenomenon is only as it appears.

The non-identity theorist would say that *tomato red* refers to the experiential phenomenon Mary first becomes acquainted with when she first looks at a

tomato after her release. However opaque that appearance is, *tomato red* only refers to that appearance, not to anything that is not revealed in the appearance.

In contrast, Loar seems committed to the notion that the opaqueness of experiential phenomena implies a lack of transparency which (somehow) allows for the possibility that scientists might eventually discover that *tomato red*, a subjective, experiential phenomenon is actually identical to some objectively measurable, physical phenomena occurring in the brain.

This seems wildly implausible for several reasons.

First, scientists tend to look for and discover causal relations between the items of interest to them. Even if one can't rule out a priori the possibility that they might discover an identity, the odds would seem to be very much against it.

Second, discovering a physical/phenomenal identity would involve discovering that something that exists in an experiencer dependent way is identical to something that exists in an experiencer independent way. Loar makes no attempt to explain how this is even conceivable let alone possible. He could, of course, simply deny that anything exists in an experiencer dependent way; but, that would give his approach an eliminative character that I don't believe he wants.

Third, on those occasions where the experiential phenomenology is at least partially transparent, there is little temptation to conclude that the physical reality is nothing other than the phenomenal reality as it appears to the experiencing subject.

This last point bears amplification.

In the case of the phases of the Moon, the appearance we call the Crescent Moon does it fact present it self as being an appearance of the Moon. I know that I am looking at the Moon rather than, say, Mars or Jupiter; but, I also know that I am seeing the Crescent Moon rather than the Full Moon.

Nevertheless, despite the transparency, I am not tempted to say that the Crescent Moon is identical to the Moon itself. Intuitively and conceptually, an appearance is not identical to the reality of which it is an appearance. The Moon has multiple appearances each distinguishable from the others; so, none can be identical to the Moon itself.

Perceptual examples are possible as well. Every few years I notice that my vision has again become blurry. I can't read street signs very well and distant landscapes appear a little fuzzy. I suspect that I need new glasses. I go to see my eye doctor who writes me a new prescription.

The blurriness of my vision revealed something about a physical phenomenon (age-related changes to the lens of the eye) associated with the experiential phenomenon of blurry vision; but, there is no temptation to think that physical reality has altered to remain identical to the altered manner in which it appears to me. Blurry vision is not identical to the changes to my lens. Blurry vision was caused by the changes to my lens.

The implausibility of physical/phenomenal identities is important because such identities are presupposed by the view that Loar defends.

> Here is the view to be defended. Phenomenal concepts are recognitional concepts that pick out certain internal properties; these are physical-functional properties of the brain. (Loar, 1997, 601)

For the purpose of this reply, I'll assume the truth of the claim that phenomenal concepts are recognitional concepts; but, anyone taking the phenomenological intuition seriously would say that phenomenal concepts refer to experiential phenomena.

Mary is aware of phenomenal redness and refers to it. This is her situation phenomenologically: She is presented with and/or becomes acquainted with an experiential phenomenon previously unknown to her; and, she refers to it when she says something like, "So, *that* is tomato red!".

It would be open to the identity theorist to argue that experiential phenomena are identical to physical phenomena; and, therefore, that whenever one refers to an experiential phenomenon one is automatically referring to the physical phenomenon to which it is identical – whether one wants to refer to physical phenomena or not.

However, Loar presents no such argument. He simply assumes that such identities are true. Speaking in the language of properties, he says that, according to physicalist, "experiential properties are physical properties" (Loar, 1997, 599). I take it that his commitment to physical/phenomenal identity claims would survive translation into the language of phenomena.

> They are the concepts we deploy in our phenomenological reflections; and there is no good philosophical reason to deny that, odd though it may sound, the properties these conceptions *phenomenologically reveal* are physical-functional properties - but not of course under physical-functional descriptions. (Loar, 1997, 601-602)

Now I would readily agree that I am presented with phenomenology; but, since phenomenology is usually opaque, it usually doesn't reveal anything. Nevertheless, according to Loar, phenomenal concepts phenomenologically reveal physical-functional properties.

I'm concerned about the phenomenological revelation itself. Given that the phenomenological revelation is not identical to the property it reveals, I should be able to refer to various aspects of the revelation without thereby referring to any brain activity at all. However, according to Loar (and other PCS tacticians) phenomenal concepts refer to physical properties of the brain.

If terms like *phenomenal redness* and *tomato red* are defined to refer to physical phenomena rather than to the phenomenological revelation of the same, we would need another set of terms to refer to aspects of the phenomenological revelation itself.

Is the phenomenological revelation as it is experienced by me identical to the physical properties or physical phenomena of which it is a revelation? If so, Loar would need to explain how the subjective revelation facilitated by phenomenal concepts could be identical to some physical phenomenon that exists objectively in the brain. If not, it would seem that Loar has reinvented the appearance/reality distinction under a new name, the property/revelation

distinction.[29]

* * *

Loar's execution of the PCS has serious flaws.

In common with most other PCS tacticians, Loar does not present a case for the identity claim upon which his argument depends. I understand that he is standing his ground against anti-physicalist arguments and intuitions; and, he may be completely justified in holding that he's not required to actually argue for the identity claim on which he bases his defense. But, unless someone somewhere defends the physical/phenomenal identities essential to identity theory physicalism, there is the problem of resting the defense of physicalism on a potentially vacuous conditional claim.

Loar makes a point of trying to take the antiphysicalists' arguments and intuitions as offered before defending against them; and, can hardly be faulted for defending against an argument that assumes that tomato red is some sort of property. Some antiphysicalists do present their argument that way.

However, I share with Churchland the belief that the key assumption the KA challenges is the assumption that there is only one kind of phenomenon; so, I've recast the antiphysicalists' challenge in terms of phenomena; and, I've shown that identity theory physicalism faces an insurmountable problem in the appearance/reality distinction. Unless the identity theory physicalist can show that something objectively detectable/measurable in the brain is nothing other than an appearance to a subject, our choices reduce to the choice between eliminative materialism and some form of non-identity theory.

It can easily be argued that Loar failed in his stated intention of taking the antiphysicalists' intuitions at they are presented. A number of philosophers (not all of them physicalists) hold some version of the thesis of Revelation; but, Loar doesn't address this issue. He simply assumes that, although appearances are opaque, one can somehow pierce the veil of appearance and discover the true, physical reality which the appearance is an appearance – or *revelation*.

While that assumption may not be controversial by itself, it certainly becomes controversial when combined with a thesis about the language of discourse; namely, that the terminology by which one would attempt to refer to an appearance or revelation actually refers to the physical property of which the appearance is an appearance or of which the revelation is a revelation. That thesis presupposes the identity of the revelation and the item revealed, making Loar's argument question begging.

Finally, Loar introduces the concept of phenomenological revelation without clarifying its relation to the other items he discusses. To this observer, it sounds like Loar has simply reinvented the appearance/reality distinction using new terminology. If so, it undermines his own argument. Just as an appearance is not identical to the reality of which is is merely an appearance, one naturally assumes that the phenomenological revelation of a physical-functional property

---

29 I am not alone in thinking that Loar's account of phenomenal concepts recreates the appearance/reality distinction. See Balog (2012, 25) for a similar observation.

is not identical to the property of which it is a phenomenological revelation.

In the case of Mary, one might well say that the experiential phenomenon, tomato red, she becomes acquainted with upon her release from monochromatic confinement is the phenomenological revelation of a physical phenomenon, NFP-4738, she already knew about. But, unless the revealing phenomenon, *tomato red*, is identical to the revealed phenomenon, NFP-4738 (or some quantum microtubular computation or some physico-chemical property of the brain or whatever), we have two distinct items, one physical (or physical-functional) and one experiential. In which case, Loar has failed to save identity theory physicalism from a weakened form of the knowledge argument, KA:TNG.

### §3.3.3.6.3.3 Reply to Balog

Katalin Balog (2012) a physicalistic account of phenomenal concepts which is intended to provide an account of acquaintance while being neutral as between "physicalist and dualist accounts of qualia in that both metaphysical views are compatible with it" (p. 12)

She has given a description of acquaintance that is indeed quite reasonable; and, my notion of acquaintance is highly similar, with one notable difference.

According to Balog, acquaintance is a unique epistemic relation between a person and his or her "phenomenally conscious states and processes" which provides direct and incorrigible access to those states "in a way that seems to reveal their essence".

I have reservations concerning the term "state".

We expect uses of "state" to specify *that* of which the state is a state. Suppose that "phenomenal state" referred to a physical state of the brain that we think is also a phenomenal state because there is some experiential phenomena associated with it.

In my view, I am acquainted with the experiential phenomenon only. I am not acquainted with neural firing patterns, quantum microtubular computations or any other neural (physical) phenomena that differentiates one brain state from another.

In contrast to Churchland who clearly believes that the experiencing subject is directly acquainted with physical phenomena[30], it's reasonably clear that Balog believes that acquaintance with a phenomenal state is acquaintance with the experiential phenomena associated with that state. When addressing the nature of acquaintance, Balog says that acquaintance reveals "the core feature of phenomenal states ... their phenomenality". (2012, 30)

So, the points that needs clarification are whether an experiential phenomenon such as phenomenal redness is itself a state and, if so, of what is it a state.

Balog's account, which she sometimes calls the quotational account and at other times she call the constitutional account,

---

30 I deny having the power of philosophical clairvoyance, the power to become directly acquainted with an external physical reality.

> ... does not explain the phenomenality of brain states – it accepts and *explains* the existence of an explanatory gap between phenomenal and physical descriptions. (2012, 31)

Her objective is to show that the features of the explanatory gap arise from "the special cognitive architecture involved in phenomenal concepts", which architecture

> ... is neutral with respect to metaphysical nature of the phenomenal states involved. It is thus open to the physicalist to maintain that types of brain state are identical with types of phenomenal state. (2012, 31)

Presumably, it is also open to the dualist to maintain that the experiential phenomenon that makes a brain state a phenomenal state of the brain is not identical to the physical phenomenon that differentiates one physical state of the brain from another.

The constitutional account would work in each case.

> The constitutional account proposes that a certain kind of concept refers to something that (partly) constitutes it, and refers to it in virtue of it being so constituted but no actual account has been proposed of how a concept can be like that. How can constitution determine reference? (Balog, 2012, 32)

> I will try to make it plausible that, in the particular case of direct phenomenal concepts, reference is determined by constitution. (Balog, 2012, 32-33)

Balog's theory is that a phenomenal concept is *quotational*, it incorporates an token of the experiential phenomenon to which it refers and by which it is constituted.

Overall, it's an ingenious theory, one that deserves further attention from philosophers of all persuasions; but, it also has a flaw that may not be repairable: dualists don't need phenomenal concepts, just terms that may be used to refer to experiential phenomena.

It is widely held that Mary can be proficient in the use of color terms if she has enough information about the circumstances in which a term is used. So, upon becoming acquainted with the experiential color phenomenon, *tomato red*, Mary finally becomes acquainted with the *referent* of the term which she knew all along and which she could use correctly in some circumstances.

Thus, in my admittedly dualistic view, a reference fixing event links the term and the experiential phenomenon to which it refers in the experiencing of a particular subject. If that is enough to constitute a phenomenal concept, then dualists would have phenomenal concepts; but, it's not clear that such a weak version of the phenomenal concept concept can help physicalists. Phenomenal concepts may have to be more peculiar when employed by physicalists; not having a peculiar kind of phenomena, they may require a peculiar kind of phenomenal concept. Time will tell.

That said, however, I must also say that I don't see how Balog's approach helps physicalists defend physicalism from the various anti-physicalist arguments Balog is concerned with. Thus, while the identity theory physicalist may "maintain that types of brain state are identical with types of phenomenal state",

> ... there is no explanation of why this brain state type (neurophysiologically or functionally characterized) is identical with a phenomenal state type (phenomenally characterized) – hence the explanatory gap – but there is an explanation in terms of the constitutional account of why there is an explanatory gap even if physicalism is true. (Balog, 2012, 31)

Let us assume arguendo that the following conditional is true.

> If physicalism is true, the constitutional account explains why an explanatory gap remains.

Without some showing that physicalism is actually true, that conditional is potentially vacuous. If physicalism is false, the conditional becomes vacuously true; but, the constitutional account would still (allegedly) explain why there would be an explanatory gap *if* physicalism were true.

Who would abandon the powerful intuition of distinctness because of a potentially vacuous conditional claim?

Balog herself may not be personally obligated to provide an affirmative defense of the physical/phenomenal identity claims upon which identity theory physicalism depends. There may be a division of labor in philosophy as in other disciplines. But, until *someone* actually does provide the missing argument for physical/phenomenal identities, there is a gap in the defense of identity theory physicalism that is being plugged by nothing more than a potentially vacuous conditional.

That doesn't sound like a strategic victory for physicalism to me. Indeed, mysterians would likely regard the success of Balog's analysis as a strategic victory for mysterianism.

### §3.3.3.6.4 Claiming an A Posteriori Identity

Under the influence of Kripke, Putnam and others, many philosophers have come to believe that some identities – *a posteriori identities* – can only be known on the basis of experience. The paradigmatic example of an a posteriori identity is ...

> **[WH] Water is $H_2O$**

By itself, [WH] is irrelevant to both Jackson's KA and my KA:TNG. However, the possibility that there are *any* a posteriori identities at all appears to offer new hope to identity theory physicalists. Instead of being required to produce *a priori* arguments for physical/phenomenal identities, *a posteriori* physicalists argue that such identities can only be known on the basis of experience – empirical inquiry. Consequently, the argument is that we must wait for scientists to decide whether they are true or false.

This promissory identity theory is the essence of the a posteriori physicalist's reply to Jackson's KA.

*Is there a similar reply to KA:TNG?*

As pointed out above, KA:TNG targets identity theory physicalism. It's weaker than Jackson's KA in the sense that it makes not attempt to refute versions of physicalism that are compatible with having two fundamentally distinct kinds of

phenomena, experiential and physical$_1$.

On the other hand, being weaker than the KA, it is also more robust (less fragile). The only way to defeat KA:TNG is to prove phenomenon monism: either by showing that there are no experiential phenomena at all or by showing that, for each experiential phenomenon, there is a physical$_1$ phenomenon to which it is identical. Failing that, the opponent of KA:TNG may aim for a stalemate by showing that a proponent has failed to prove that there are experiential phenomena distinct from any physical$_1$ phenomenon.

In my view, scientists are unlikely to meet this criteria because it would involve showing that some physical$_1$ reality is nothing other than an appearance to an experiencing subject; thereby collapsing the appearance/reality distinction for experiential phenomena such as color phenomena and the physical reality of which those appearances are merely … *appearances*.

There is no evidence that any *scientists* are actually working to show that an appearance to an experiencing subject can be identical to the physical$_1$ reality of which it is an appearance or to any other physical$_1$ reality with which it may be associated. So, a question naturally arises as to the basis for the a posteriori physicalist's hope that scientists will someday deliver them from allegations of physical/phenomenal non-identities and, more generally, from the non-identity theory of the brain/experience relation.

*What exactly are a posteriori physicalists expecting from scientists?*

It is commonly thought that identities are necessary rather than contingent, if true at all. Yet there is a sense of contingency about [WH]. People have always known about water; but, we only recently learned about chemistry. There was a time, shortly after modern chemistry began, during which it was reasonable to suspect that water had a chemical composition and to entertain conflicting theories as to what that composition might turn out to be. In actual fact, it turned out that water is predominantly composed of $H_2O$; but, there was a time during which it may have been thought that water might turn out to be predominantly composed something else.

This sense of contingency carried over to philosophical thinking about alleged physical/phenomenal identities; and, indeed, they were once considered contingent identities.[31]

> There did seem to be a tendency among philosophers to have thought that identity statements needed to be necessary and a priori truths. However identity theorists have treated ‘sensations are brain processes’ as contingent. We had to find out that the identity holds. (Smart, 2007)

More recently, philosophers have come to think that the sense of contingency is an illusion deriving from a failure to appreciate an implication of having to find out that the identity holds; namely, that not all identity claims are knowable a priori.

---

31 Some philosophers continue to defend the concept of contingent identity. See, for example, Gibbard (1975). However, an examination of the examples offered reveals that they are the same examples offered to show constitution rather than identity.

However, holding that a posteriori identity statements are necessarily true, if true at all, has consequences. It is generally held by those who argue for such identities that they are metaphysically necessary; meaning, true in all possible worlds, if true at all; so, here, philosophers may have overreached themselves.

In my view, it is unrealistic to expect scientists to discover metaphysically necessary identities simply because scientists can't tell us what is true in all possible worlds. At best, they can only tell us what is true because of the physical laws of this world (nomological necessity) and what is true only because of the way scientists have defined their terms (analytic necessity).

It is more plausible to suppose that a posteriori physicalists hope that scientists will eventually discover empirical evidence on the basis of which philosophers will successfully argue that pain is identical to firing C-fibers; that tomato red is identical to neural firing pattern NFP-4738 (or quantum microtubular computation QMC-8113 or whatever); and, similarly for all other claims of physical/phenomenal identity.

It may be that the epistemological journey is like a relay race in which scientists carry the baton as far as they can before passing it off to philosophers who proceed from there.

> … the identification of a neural feature that correlated perfectly with consciousness would still leave open a certain metaphysical question: is the relation between consciousness and the relevant neural feature merely correlation, or is that correlation indicative of a deeper, more intimate relation between the two? Work addressing this further question can be thought of as attempting a philosophical interpretation of scientific theories, somewhat on a par, say, with philosophical interpretations of quantum mechanics: in both cases, philosophy has to take over where science proper ends in order to articulate an intelligible conception of how the world must be given what the science suggests. (Kozuch and Kriegel, 2015, 401)

Unfortunately, if philosophy begins where science ends, one must wonder about the consequences of allowing philosophers to assume their way across the finish line – for they have no other way to proceed once empirical inquiry has reached its limits.

In the case of [WH], which a posteriori physicalists believe has already been proven, any assumptions philosophers add to the empirical facts to reach [WH] may undermine their attempts to claim that [WH]

1. is true;
2. is metaphysically necessary, if true;
3. is an a posteriori truth; and,
4. is an identity statement.

In what follows I will assume that [WH] is true in the sense that there is some reading of [WH] such that it is (or, at least, appears to be) true on that reading. However, I will not assume that the identity reading is the reading that is true. [WH] is ambiguous in that it is not clear what sense of *is* is used in asserting [WH].

Reading [WH] as using the is of identity yields the identity reading of [WH]

favored by a posteriori physicalists:

**[WH-1] Water is identical to $H_2O$**

However, there are those who argue that [WH] is true only when read as using the is of constitution, yielding the constitutional reading:

**[WH-2] Water is constituted by $H_2O$**

It is also possible to read [WH] or something close to it as using the is of predication or class inclusion, yielding:

**[WH-3] $H_2O$ is an instance of water**

It seems inconceivable that [WH-3] could be false. Whether we start with a sample of lake water, sea water, well water, bottled water or even non-potable water, once we remove from our sample everything that isn't $H_2O$, what we are left with – pure $H_2O$ – is still water.

What we couldn't say is that what's remains after this purification process is still water, the natural kind (henceforth, $water_{nk}$). Surely, odd though it may seem at first glance, artificially purified water – $H_2O$ as pure as technologically possible – is not an instance of $water_{nk}$.

So, while [WH-3] is certainly true, it is using "water" in its most inclusive sense, as the name of the class consisting of all instances of $water_{nk}$ plus artificially purified water, pure $H_2O$.

[WH], on the other hand, is thought to be a theoretical identity; meaning, an identity statement linking a term for a natural kind and a term for a scientific kind; so, it seems reasonable to assume that, in [WH] "water" should be taken to mean $water_{nk}$.

In considering the truth of [WH-1] and [WH-2], we will need to guard against a similar shift in meaning. They should each be statements about $water_{nk}$; otherwise, we've already lost our connection to our stating point, the claim that [WH] is a theoretical identity known a posteriori.

This point should also be kept in mind when considering what I will call the *essentialist reading* of [WH]:

**[WH-4] Water is essentially $H_2O$**

Leaving aside for the moment any concern about what may be true of other possible worlds, I doubt that anyone would contest the claim that, in this actual world, containing $H_2O$ is essential for being $water_{nk}$. But, when we read [WH-4] as being about $water_{nk}$, it is not clear whether to understand it as

**[WH-4.1] $Water_{nk}$ is essentially $H_2O$ and nothing else**

or

**[WH-4.2] $Water_{nk}$ is essentially $H_2O$ plus impurities**

The former seems reductive in a way that the latter does not.

As we shall see, much of the debate revolves around the question of whether $water_{nk}$ is *reducible* to being nothing more than its essential ingredient. In the

course of the debate over the virtues of an essentialist reading of [WH], we would need to guard against the practice of starting out to show with $water_{nk}$ is identical to a scientific kind, $H_2O$, reducing $water_{nk}$ to the essence of water and then alleging an identity between $H_2O$ and the essence of water. As an attempt to demonstrate the theoretical identity between a natural kind and a scientific kind, such a practice perpetrates a bait and switch operation.

### §3.3.3.6.4.1 Scientific Essentialism and KA:TNG

George Bealer (1994) argues that the principles of scientific essentialism, a currently popular view developed by Kripke and Putnam, gives the a posteriori physicalist a reply sufficiently strong to defeat Jackson's KA. Bealer goes on to argue that those opposed to identity theory physicalism may rely on a strengthened version of the certainty argument. It's an interesting argument; but, I will take no position concerning its soundness. My concern is with defending KA:TNG against the reply that, according to Bealer, succeeds against Jackson's KA.

Bealer's argument seems to be that a posteriori physicalists may reasonably hope that scientists will someday *define* experiential phenomena to be identical to $physical_1$ phenomena just as they have already defined water as being identical to $H_2O$.

> Identity theorists who accept SE [Scientific Essentialism] should diagnose the situation thus. Being in pain is in fact having firing C-fibers. Yet I can be certain that I am in pain and not certain that I have firing C-fibers. The reason is that I do not know the relevant definition: x is in pain $iff_{def}$ x has firing C-fibers. This definition is a scientific definition (akin to the scientific definition of water: x is water $iff_{def}$ x is $H_2O$) (Bealer, 1994, 191)

This seems a bit far-fetched, though.

In my view a statement such as “x is in pain iff x has firing C-fibers” is a statement about an empirically discovered correlation rather than a definition because it specifies the circumstances in which a given experiential phenomenon occurs in humans. But, assuming that we are somehow justified in taking it as a definition, we can easily give Mary analogous definitions before she is released; for example,

> **[NAPI-1] Subject s is experiencing *tomato red* if and only if the brain of subject s exhibits neural firing pattern NFP-4738.**

Or instead, perhaps,

> **[NAPI-2] Subject s is experiencing *tomato red* if and only if the brain of subject s exhibits quantum microtubular computation QMC-8118.**

Suppose the truth turned out to be [NAPI-1] and we presented that definition to Mary via her monochromatic television monitor. She now has complete knowledge by description of the definition. She already had complete knowledge by description of one relevant phenomenon, NFP-4738.

Bealer's argument seems to be that a posteriori physicalists may claim that Mary doesn't know the definition, at least not fully, until after she experiences *tomato*

*red* for herself. Indeed, he imagines the identity theorist admitting that the definition is not itself a physical fact; rather, it is a *definitional* fact.

> … if Mary is asked to tell us the definitional facts after her release, she would give exactly the same answers she would give if she were asked before her release. Nor would her words have changed their meanings; she would just understand them more fully. With this in mind, identity theorists should put their point thus: anyone who knows with understanding all the physical and definitional facts is in a position to know with understanding all the facts. (Bealer, 1994, 192, fn 9)

In reply, non-identity theorists would say that the physical facts are facts about physical phenomena that Mary knew about before her release. The so called definitional facts that Mary can only learn from experiencing are facts about the experiential phenomena with which she becomes acquainted after her release.

Thus, the reply that Bealer says is available to a posteriori physicalists effectively concedes that there are two kinds of phenomena. It doesn't matter whether we call them physical and experiential or (with the subjective physicalist) objective physical and subjective physical or (with Bealer) physical and definitional or something else.

It may be argued that everything discussed up to this point is consistent with the possibility that, for each experiential phenomena, there is a physical$_1$ phenomenon to which it is identical. Bealer argues along these lines but using the language of properties:

> Because her ignorance of the definition is entirely consistent with the thesis that experiencing red is identical to a physical property, the knowledge argument is no threat to the identity thesis. (Bealer, 1994, 191)

I've already given my reasons for thinking that this will never happen; but, those reasons assume that it is appropriate to conduct the debate in terms of phenomena rather than properties.

*How does the debate turn on this terminological difference?*

If phenomenal redness and similar items of interest in the debate of KA:TNG are phenomena rather than properties, we may credit Mary with knowledge sufficient to conclude that an experiential phenomenon is not identical to a physical phenomenon.

If Mary classifies *tomato red* as an experiential phenomenon, we may credit Mary with knowing that it is an appearance to an experiencing subject. From her lessons and her independently conducted empirical investigations, we may credit Mary with knowing that the physical universe exists independently of her existence and independently of her experience of it. Allowing her these bits of knowledge gives Mary sufficient reason for concluding that the physical phenomena she already knows about from her lessons are something other than mere appearances to an experiencing subject.

Indeed, we can reasonably credit Mary with knowing, *before her release*, about the appearance/reality distinction with respect to other sense modalities. So, if Mary knows something like [NAPI-1] before her release, one may plausibly argue that Mary also knew, prior to her release, that “tomato red” refers to an

appearance, an experiential phenomenon with which she was not yet acquainted.

After her release she becomes acquainted with that experiential phenomenon. She then knows what the term “tomato red” refers to. But, even prior to her release, she already knew that it would be an experiential phenomenon to which no physical phenomenon could be identical.

*Could we deny that Mary has knowledge of the appearance/reality distinction?*

It seems odd that, despite being a brilliant student, a super-scientist and an ideal reasoner, Mary is somehow ignorant of the appearance/reality distinction despite having experience in other sense modalities besides color vision.

*Could we deny that Mary has knowledge of the ontological status of the physical world?*

We could assume that Mary is locked in a state of Cartesian style hyperbolic doubt as to the consciousness-independent existence of the physical universe; but, insisting that Mary be ignorant of facts every physicalist takes for granted seems to be an implausible way to defeat KA:TNG.

### §3.3.3.6.4.2 The Flagship Identity Claim

Bealer tells us that arguments supporting Scientific Essentialism rely on intuitions supported by thought experiments such as the Twin-Earth gedanken. Intuitively, it is said, if we went to Twin-Earth and found stuff that was macroscopically like water in all respects but turned out to be composed of XYZ rather than $H_2O$, it would not be water.

> Suppose that this and kindred intuitions are correct, and suppose that all and only samples of water are as described. Then, we may conclude that, in all actual and counterfactual situations, *something would be composed of water if and only if it were composed predominantly of $H_2O$.* In turn, we may conclude that, **necessarily, water = $H_2O$.** (Bealer, 1994, 192 (emphasis supplied))

From this passage, I take the two assumptions of scientific essentialism (as applied to the water/H2O relation) to be as follows:

> **[W-1] Something is water if and only if it is composed predominantly of $H_2O$ in all actual and counterfactual situations.**
>
> **[W-2] (therefore) Necessarily, Water is $H_2O$**

To facilitate clarity, I've translated [W-1] from the subjunctive to the indicative mood; and, I've interpolated '(therefore)' into [W-2] to make it clear that we are supposed to be able to deduce it from [W-1] in some unspecified way. For [W-2], I've translated the '=' sign into 'is' to embed [WH], our flagship example of a necessary a posteriori identity, within it.

Now here is where I part company with Bealer. He seems to think that, if the principles of Scientific Essentialism generalize from cases like [WH] to cases of alleged physical/phenomenal identities, the KA will fail. He later argues that allegations of physical/phenomenal identities fail his more revised certainty argument. My reply is that Scientific Essentialism fails even for cases like [WH]; so, there is no reason to think it will succeed against KA:TNG.

Note that [W-1] is presented as a biconditional to give the necessary and sufficient conditions for being water. So, in principle, there may be challenges to each of these conditions, to the deduction of [W-2] from [W-1], to the allegation of necessity and to the interpretation of the embedded [WH] as an identity claim rather than, say, a claim of constitution. I'll now consider some of these challenges.

### §3.3.3.6.4.2.1 The Deduction of Absurdity

My main objection to the philosophical doctrine known as Scientific Essentialism concerns the move from [W-1] to [W-2].

I'll have two arguments with which to support this position. First, I'll argue that there is no direct step from [W-1] to [W-2]; so, a philosopher who wants to reach [W-2] or (or the embedded [WH]) will have to add further assumptions. I will later examine a number of arguments for and against [WH] to draw out these assumptions.

Second, I will argue that, on the identity reading of [WH], attempting to derive [W-2] from [W-1] is a fool's errand – whatever other assumptions a given philosopher adds to the mix. To see why let us consider how to translate [WH] into a statement that is explicitly an identity statement.

**[WH-1] Water is identical to $H_2O$**

This may be the obvious candidate for an explicit formulation of the identity reading of [WH]; but, it could be further expanded in various ways to clarify the meaning of identity.

**[WH-1.1] Water is nothing other than $H_2O$**

**[WH-1.2] Water is $H_2O$ alone**

**[WH-1.3] Water is $H_2O$ and nothing else**

[WH-1.1] has the virtue of capturing Frege's insight into the meaning of identity. In discussing another classic example of a necessary a posteriori identity,

**[V-1] The Morning Star is Venus**

Frege paused to explain the difference between the *is* of identity and the *is* of predication.

In identity statements, Frege tells us, “'is' is used like the 'equals' sign in arithmetic to express an equation” and, in a footnote to “equation”, Frege explained that, “I use the word 'equal' and the symbol '=' in the sense 'the same as', 'no other than', 'identical with'” (Frege, 1960, 44). He then continued:

> In the sentence 'the morning star is Venus', 'is' is obviously not the mere copula; its content is an essential part of the predicate, so that the word 'Venus' does not constitute the whole of the predicate. One might say instead: 'the morning star is no other than Venus'; what was previously implicit in the single work 'is' is here set forth in four separate words, and in 'is no other than' the word 'is' now really is the mere copula. What is predicated here is thus not Venus but no other than Venus. (Frege, 1960, 44)

However, [WH-1.1] is problematic because combining one half of the

biconditional in [W-1] with [WH-1.1] yields an absurdity:

> **[W-3] If something is composed predominantly of $H_2O$ it is water; and, therefore, it is nothing other than $H_2O$.**

Intuitively, if A is composed predominately of B, it is not that case that A is only B. Suppose we find that the Army is composed *predominately* of men. We would conclude that it is ***not*** the case that the Army is composed *only* of men. Similarly, if water is composed predominately of $H_2O$, it is not the case that water is only $H_2O$; although, significantly, something that is only $H_2O$ is still an instance of water (in the most inclusive sense).

Statements analogous to [W-3] can be constructed using [WH-1.2] or [WH-1.3] instead of [WH-1.1]. If all are rejected as absurdities, as I think they should be, we can reject all or part of [W-1] by modus tollens.

*What part(s) of [W-1] shall we reject?*

### §3.3.3.6.4.2.2 Is $H_2O$ Necessary for Water?

We can take one side of the biconditional to state a claim about what's necessary for being water.

> **[W-1.1] In all actual and counterfactual situations, if something is water it is composed predominantly of $H_2O$**

Striking the qualification concerning counterfactual situations yields

> **[W-1.1.1] In all actual situations, if something is water it is composed predominantly of $H_2O$**

It seems reasonable to assume that [W-1.1.1] is true of the actual world; but, that only makes it a statement of nomological necessity. In contrast, [W-1.1] purports to be true in all possible worlds, making it a claim of metaphysical necessity. Hence, the concern with thought experiments involving visits to Twin-Earth, the hypothetical planet postulated by Putnam.

Travelers returning from Twin-Earth report that there are lakes, rivers and oceans full of stuff that appears to be water. It looks like water, tastes like water, quenches thirst like water, falls from the sky as rain or snow and so on. However, a chemical analysis reveals that this stuff contains no $H_2O$ at all; it's a complex chemical whose formula is symbolized, XYZ.

When travelers returning from Twin-Earth are polled as to what their intuitions tell them. It is widely assumed among Earth philosophers that all returning travelers will say something like, “on Twin-Earth there is no water, only twin-water”. However, I can tell you that I've just returned from a visit to Twin-Earth and I am reporting, “on Twin-Earth, the water is composed of XYZ!”.

Strangely enough, it appears that it is not metaphysically necessary that I (or anyone) adopt the most popular intuition concerning the water on Twin-Earth. So, here I am, cogitating in total defiance of (allegedly) metaphysical necessity; and yet, I've perpetrated no logical contradiction.

My reasons for preferring this unpopular intuition are based on Bealer's (1987)

distinction between functional and compositional kinds of stuff.

There was a time, Bealer tells us, that the only available fuels were solid hydrocarbons. But, when people discovered liquid hydrocarbons they had no trouble considering them as a new kind of fuel. Consequently, we might say that fuel is a functional kind of stuff; meaning, that whatever fulfills the function of fuel is a kind of fuel.

Bealer accepts the popular intuition that travelers returning from twin-Earth will all say that there is no water there. His reason is that water is a compositional type of stuff rather than a functional kind of stuff; so, when humans from Earth find a substance on twin-Earth that functions like water in all macroscopically noticeable ways is not composed predominantly of $H_2O$, they will *inevitably* say that what functions like water on Earth isn't water because it isn't composed as water is composed on Earth.

The problem is that there seems to be no logical reason why they could not say instead that what functions like water on Earth is water although it is not composed as water is composed on Earth.

It may be argued that no one can say that the water on Twin-Earth is composed of XYZ because water just is a compositional kind of stuff; but, that would be question begging. It also seems to get the cart before the horse. Is there a fact of the matter as to whether water is a compositional kind of stuff rather than a functional kind of stuff; or, do we classify water one way or the other because of what we imagine people might say in hypothetical situations?

Interestingly enough, Bealer says that food is a functional kind of stuff. So, if Earth humans visiting Twin-Earth find that what looks like food, tastes like food and functions like food to sustain the body, they'll consider it to be food. If they find something that looks like a watermelon and tastes like a watermelon they'll consider it a watermelon, a kind of food, even if it contains XYZ instead of $H_2O$.

So, travelers returning from Twin-Earth would say that there is food but no water there, which contradicts the intuition that both food and water are necessary to sustain life. I prefer to preserve that intuition and conclude that, if our astronauts learn that the stuff on twin-Earth that functions like water isn't composed of $H_2O$, they would nevertheless report that Twin-Earth is a habitable world that has both food and water readily available in quantities sufficient to sustain human life. Of course, we'd expect footnotes in the official report indicating that the food on Twin-Earth is composed of … (whatever) and that the water on Twin-Earth is composed predominantly of XYZ.

I concede that Putnam's intuition is more popular here on actual Earth at this time. That may change, of course; but, what is the situation on other possible worlds right now? It seems rather easy to imagine a world – I call it Planet Heresy – which is just like our actual world so water is composed predominantly of $H_2O$. The only difference between Earth and Heresy is that most of the heretics returning from Twin-Earth say things like “curiously, the water on Twin-Earth is made of XYZ” or “their water is XYZ”.

The philosophical community on Heresy concludes that [W-1.1] is false while [W-1.1.1] is true on Earth and Heresy but false on Twin-Earth. While I personally

agree with the philosophers of Heresy, I take it that most contemporary Earth philosophers would hold that [W-1.1] is true universally.

It would even be true on Twin-Earth even though the philosophers there (under the sway of their Kripke, no doubt) would say that

> **[W-1.1.2] In all actual situations, if something is water it is composed predominantly of XYZ**

I don't see how we could ever, even in principle, resolve the debate between our Kripke and their Kripke. Both Kripkes would say that “water” is a rigid designator; although, they would disagree as to what it rigidly designated. Our Kripke would say that “water” refers to stuff composed predominantly of $H_2O$. Twin-Earth Kripke would say that “water” refers to stuff composed predominantly of XYZ.

How would we resolve the discrepancy between the conclusions of the philosophers of Earth and the philosophers of Heresy as to whether [W-1.1] is true? How long would the interplanetary debate have to go on before we realized that it was driven by competing philosophical intuitions?

In the event that we recognize that there are competing philosophical intuitions involved, it's not clear what we should say of the matter. Perhaps, to those (and only those) in the grip of one particular philosophical intuition, it will appear as if the claim that $H_2O$ is necessary for water is a metaphysically necessary truth. Those in the grip of a different intuition may reach a different conclusion.

In any case, the claim that $H_2O$ is necessary for water fails because it hasn't been shown that a philosopher's preferred intuition is anything more than a personal idiosyncrasy. The popularity of a philosophical intuition may explain the appearance of metaphysical necessity; but, it does not guarantee the actuality of metaphysical necessity.[32]

### §3.3.3.6.4.2.3 Is $H_2O$ Sufficient for Water?

We can construct a statement based on [W-1] concerning the sufficiency of $H_2O$ for water.

> **[W-1.2] In all actual and counterfactual situations, if something is composed predominantly of $H_2O$, it is water.**

David Barnett (2000) challenges [W-1.2] by denying that satisfying it is sufficient for being water. He defends this intuition by supposing that someone exploring Twin-Earth might discover something composed entirely of $H_2O$ molecules but which, instead of exhibiting any of the manifest qualities of water, looks like a mushroom. Intuitively, says Barnett, this would not be considered water.

In my view, Barnett's example is weak precisely because the hypothetical mushrooms lack the manifest qualities by which we identify and re-identify instances of the natural kind, water.

---

32 Perhaps we need to develop a concept of relative truth for those statements that appear true from a certain point of view – to the someone with certain philosophical intuitions.

A more plausible test of [W-1.2] would involve finding something that exhibits at least some of the manifest qualities of water and is composed predominantly of $H_2O$ but is (allegedly) not water. Fortunately, we do not have to visit a hypothetical planet to find candidates for challenging [W-1.2]. There is, here on actual Earth, a substance that challenges our willingness to accept [W-1.2]: *methane hydrate*.

Methane hydrate is ice in which methane is trapped within the crystalline lattice formed by $H_2O$ molecules hydrogen bonding with their neighbors.

*Is methane hydrate water?*

Glacial ice contains small pockets of trapped air which allows scientists to measure the amount of various gases (including carbon dioxide and methane) in the atmosphere of past ages (Bender et al., 1997). Glacial ice doesn't cease to be ice (and, therefore, water) just because it contains trapped air; so, it is not clear why ice should cease to be ice (and, therefore, water) just because it contains trapped methane.

Perhaps the difference is in the manner of the trapping. In glacial ice, the trapped gases are found within gaps in the crystalline lattice that is the ice. In methane hydrate, the trapped methane is found inside the crystalline lattice when molecules of $H_2O$ form a cage trapping a molecule of methane inside it.

To preserve [W-1.2], I'm willing to bite the bullet and say that methane hydrate is a form of water because it is composed predominantly by $H_2O$. Someone might challenge this claim by saying that there is too much methane in methane hydrate to say that a chunk of methane hydrate is predominantly composed of $H_2O$. However, any attempt to make the question of what counts as water depend on the percentage of $H_2O$ it contains will run into a problem noted by Joseph LaPorte (1998). Water from Utah's Great Salt Lake contains less $H_2O$ than do chickens, baby humans and tomatoes. Yet, the stuff in the Great Salt Lake is water but chickens, human babies and tomatoes are not.

Any attempt to gerrymander a definition of “predominantly” so that it includes stuff that we think should be called water (eg. lake water, distilled water, mineral water, alkaline water (pH = 9.5!), ozonated water, non-potable water, etc.) and excludes stuff that we think should not be called water (eg. tea, tears, chickens, babies, tomatoes, etc.) will seem arbitrary rather than necessary.

A better reply would be to adopt the view espoused by Malt (1994) that the proportion of $H_2O$ is not the sole determinant of whether a substance is or is not water. Human interests may also be a factor; but, then it would be unclear whether human interests favored considering methane hydrate to be water or non-water. Could a belief as to whether methane hydrate counts as water that is based on pragmatic human interests be a metaphysical necessity in all possible worlds?

In favor of considering methane hydrate as water is the fact that in science news items directed at the general public, methane hydrate is often referred to as *ice that burns*. That term is potentially misleading because it doesn't refer to frozen

methane. The ice that burns is water ice not methane ice.[33] But, of course, it is only the methane that burns.

However, methane hydrate is not, to my knowledge, referred to as water or as a form of water in the scientific literature. There is even one curious press release from the American Chemical Society “informing” us that

> Government researchers are reporting that these so-called "gas hydrates," a frozen form of natural gas that bursts into flames at the touch of a match, show increasing promise as an abundant, untapped source of clean, sustainable energy. (American Chemical Society, 2009)

Perhaps we should infer that there is no fact of the matter as to whether methane hydrate is or is not a form of water. This view would find support in Ned Block's view that a

> … type of indeterminacy is exemplified in the question whether $H_2O$ made out of heavy hydrogen (that is, $D_2O$) is a kind of *water* or not? There is no determinate answer, for our practice does not determine every decision about how the boundaries of a natural kind should be drawn. To decide the question of whether $D_2O$ is a kind of water, we could either decide that water is a wide natural kind in which case the answer is yes or we could decide that water is a narrow natural kind in which case the answer is no. The issue would be settled. (Block, 2003, 42)

Maybe water is an extra-wide natural kind that includes methane hydrate; or, maybe it's not. If there is no fact of the matter on this point, would it be metaphysically necessary in all possible worlds that the answer be whatever it turns out to be in this world? Probably not. It is more plausible to conclude that it would be a contingent fact about a possible world that the matter was settled one way or the other; particularly, if the answer we settle upon seems arbitrarily chosen. But if *is* a contingent fact about a world that this question was settled one way or the other, what happens to the theory of rigid designation? Supposedly, “water” designates the same stuff in all possible worlds; so, any indeterminacy as to whether deuterium oxide and/or methane hydrate are among the referents of “water” in the actual and all other possible worlds suggests that the theory of rigid designation fails with respect to water.

Whatever the fate of the theory of rigid designation, I would argue that there is no *principled and uncontentious* reason for saying that liquid water with dissolved gases is water; but, that frozen water with trapped methane or other gases is not water.

Consider the details of one proposal for extracting the methane from methane hydrate deposits. The same press release from the American Chemical Society quoted above also mentions that the process of pumping carbon dioxide gas into methane hydrate. The carbon dioxide displaces the methane which is then collected for use as a fuel.

So now we have frozen water with trapped carbon dioxide. Is there a principled and uncontentious reason for saying that liquid water with dissolved carbon

---

33 Methane freezes at -296.4 degrees F (Wikipedia, 2017-01-05); but, the temperature at the bottom of the ocean is only “between 0-3 degrees Celsius (32-37.5 degrees Fahrenheit)!” (NESTA, 2010).

dioxide is carbonated water; but, that frozen water with trapped carbon dioxide is not water of any kind? I can't think of one. It's a close question, to be sure; but, it is one that has consequences. If frozen water with trapped carbon dioxide is water, it may become more reasonable to hold that frozen water with trapped methane is also water.

There is probably much more to be said about whether methane hydrate is or is not water; but, if scientists issue a press release reporting their decision that "water" in its most inclusive sense does not include methane hydrate, I'm sure I'll learn to live with it.[34] However, I doubt that many would say that scientists simply recognized a metaphysical necessity. I'd be inclined to say it was an arbitrary decision that could have gone the other way.

In any case, Barnett would acquire a further argument that $H_2O$ is not sufficient for being water. The situation with hydrogen oxide would become more similar to the situation with silicon dioxide, $SiO_2$.

> … consider $SiO_2$, the molecular constituent of sand, quartz, and glass. ... While it may be true that quartz is necessarily composed of $SiO_2$, it is not the case that $SiO_2$ necessarily forms quartz. If both 'Quartz' and '$SiO_2$' are rigid designators, then they do *not* rigidly designate the same thing. (Barnett, 2000, 102)

Barnett concludes, "The moral of the story is that 'water is $H_2O$' does not express a necessary identity" (2000, 102).

I concur.

### §3.3.3.6.4.3 The Type Identity Reading

Some identity theorists make it clear which assumptions they add to the results of empirical research to reach an identity claim.

> There are many places in the world where we find (typically liquid) stuff that we refer to as 'water'. Suppose that we collect numerous samples of this stuff and have them analyzed in a chemical laboratory; suppose further that in all cases, we obtain the same result: apart from insignificant impurities, all samples consist of $H_2O$. In my view, this would completely suffice for a justification of [Water = $H_2O$] – provided that we accept the two basic assumptions that 'water' is a natural kind term that refers to a chemical substance, and that chemical substances are individuated by their molecular structure. (Beckermann, 2012, 71-72)

Each of these "basic assumptions" is highly problematic. While most would agree that "water" is a natural kind term, people used it for thousands of years before we learned anything about chemical substances. It refers to a kind of "stuff" that we find and identify by observation. Once humans learned about chemistry it may have been thought that water would turn out to be a *simple* chemical

---

34 Actually, they may have done that already, albeit in an indirect way. According to the International Union of Pure and Applied Chemistry (IUPAC), methane hydrate is an inclusion compound because the guest molecule (methane) is "is in a cage formed by the host molecule or by a lattice of host molecules" (IUPAC, 1997). But, if that is enough to justify denying that methane hydrate is an instance of water, there would be questions concerning other host molecules. Does a buckyball cease to be a buckyball when it has a guest molecule of something else trapped inside it? If so, what does it turn into?

substance; but, this did not happen.

> ... most philosophers are under the impression that water is nothing more than a bunch of $H_2O$ molecules. This is simply not the case. An individual molecule of $H_2O$ doesn't have any of the observable properties we associate with water. A glass of water, pure as water can be, is better understood as containing $H_2O$, $OH^-$, $H_3O^+$ and other related but less common ions, and even this is a vast oversimplification (if we could get truly pure water, which we cannot). (VandeWall, 2007, 910)

In addition to containing the dissociation products, fragments of $H_2O$ molecules, a sample of water will have association products, clusters of $H_2O$ molecules. In liquid water, $H_2O$ molecules

> … cluster to form hydrogen-bonded bicyclo-octamers $(H_2O)_8$ that can link together into larger structures. … Evidence is mounting that water in living systems naturally gathers into frameworks of 14, 17, 21, 196, 280, or more molecules. ... And support is growing behind the idea that these intricate structures play key roles in operations ranging from molecular binding to turning on and off basic cell processes. (Daviss, 2004)

Beckermann's second basic assumption is also highly dubious; indeed, in my view it is simply false. We distinguish protium oxide and deuterium oxide because they have different properties; but, they have the same chemical formula and the same molecular structure. They are both instances of hydrogen oxide. Assuming that they are the same chemical substance, they are individuated not by their molecular structure but by their atomic structure.

Of course, one might argue that protium oxide and deuterium oxide are different chemical substances because they have different properties; but, there is a price to be paid for that solution. We would have to sacrifice a principle of semantic externalism favored by some physicalist philosophers, the principle that Michael Weisberg calls the *coordination principle*: "the thesis that scientific kinds and the natural kinds recognized by natural language users line up or can be mapped onto one another one-to-one" (Weisberg, 2003, 337).

Weisberg is willing to reject the coordination principle and appeals to the practices of chemists to justify an alternate principle of individuation for chemical substances: "Chemical kinds are to be individuated with respect to structure and reactivity at the molar, molecular, and atomic levels" (2003, 339). Using this principle of individuation, protium oxide and deuterium oxide would be distinct chemical kinds.

Clearly, if a sample of $H_2O$ is itself a mixture of chemical kinds, a sample of natural kind water, $water_{nk}$, is also a mixture. The next question is whether $water_{nk}$ contains anything besides the mixture that $H_2O$ is.

Beckermann continues ...

> Would this justification of [Water = $H_2O$] be diminished if it proved impossible to explain all (but only some or even none of) the superficial qualities of water by the chemical theory of water? Would this justification be somehow less conclusive if it turned out that some of these properties are due to impurities that simply happen to be ubiquitous – just as the yellow color of gold results from the fact that almost all samples of gold contain some amount of copper? Would we then say that water is not $H_2O$ at all but

> rather $H_2O$ + a small amount of ABC? No, we would not say this. (Beckermann, 2012, 71-72)

Yes, I would say *precisely* that. $Water_{nk}$ is a mixture of $H_2O$ plus the additional substances required to account for its properties.

$Water_{nk}$ is the liquid stuff that accumulates in lakes, flows in rivers, falls as rain and so on. It's something that can be identified by ostension in the absence of scientific knowledge about its composition. A chemical analysis of various samples of $water_{nk}$ would indeed reveal that they all contain mostly hydrogen oxide, $H_2O$, with small amounts of various other substances, just as Beckermann expects them to.

However, these other substances are not always *insignificant* impurities. Sometimes they are desirable ingredients that give $water_{nk}$ properties that hydrogen oxide alone does not have. For example, $water_{nk}$ usually contains enough dissolved oxygen to allow fish to live in it. Fish would have nothing to breathe in a pail of hydrogen oxide fresh from the laboratory. The oxygen dissolved in $water_{nk}$ (not the oxygen in the $H_2O$ molecules) is what makes it true that fish can live in $water_{nk}$.

$Water_{nk}$ is a good conductor of electricity due to the presence of ions of dissolved minerals but “pure water is a poor conductor of electricity but is not a perfect insulator as it always contains ions due to self-dissociation” (Chaplin, 2015).

We may conclude from this that samples of $water_{nk}$ and samples of hydrogen oxide have different properties. Consequently, they can't possibly be identical.[35]

More formally, I deny that $water_{nk}$ is identical to $H_2O$ alone based on the following Purified Water Argument for non-identity.

> [PWA-1] $Water_{nk}$ is the stuff that fills lakes, flows in rivers and falls as rain
>
> [PWA-2] The stuff that fills lakes, flows in rivers and falls as rain is not purified water
>
> [PWA-3] (Therefore) $water_{nk}$ is not purified water
>
> [PWA-4] Purified water is identical to $H_2O$ alone
>
> [PWA-5] (Therefore) $Water_{nk}$ is not identical to $H_2O$ alone

[PWA-1] simply formalizes the functional definition of water as a natural kind. I take [PWA-2] to be uncontroversially true; so, [PWA-3] clearly follows. If we took a sample of $water_{nk}$ from a lake or river and purified it by distillation or some other process, we would *end up* with purified water; but, refuse to assume that the stuff that fills lakes, flows in rivers and falls as rain is already purified water.

*Lake water isn't just sitting there already purified!*

Now, one may plausibly say that artificially purified water *is* $H_2O$ alone; *provided*,

---

35 At the moment, I'm only considering the properties of $water_{nk}$. If we considered various kinds of prepared waters, we'd have to concede that ingredients other than $H_2O$ are also essential to sub-kinds of water. Water doesn't have to have ozone in it to be water; but, containing ozone is essential to being ozonated water. A high pH value is not essential to water; but, it is essential to being alkaline water.

that we are willing to overlook: (1) isotopic variations in the composition of hydrogen oxide; (2) the presence of ions ($HO^-$ and $H_3O^+$) into which a very small portion of a sample of $H_2O$ naturally dissociates; and, (3) the presence of water clusters such as $(H_2O)_8$ and other variations.

Someone could object to the first provision and say that only protium oxide ($^1H_2O$) is hydrogen oxide; but, saying that neither heavy water (deuterium oxide, $^2H_2O$) nor tritiated or “super-heavy” water (tritium oxide, $^3H_2O$) is an instance of water seems arbitrary. Given that protium, deuterium and tritium are all forms of hydrogen, it is more natural to say that protium oxide, deuterium oxide and tritium oxide are all forms of hydrogen oxide, $H_2O$. That means that hydrogen oxide is itself a mixture; and, therefore, that water is a mixture. The only question is whether the mixture that is $water_{nk}$ contains anything other than the mixture that is $H_2O$.

Someone might object to the second provision and say that the self-ionization property of hydrogen oxide shows that pure $H_2O$ doesn't exist in nature; but, it is hard to see how this helps show that $water_{nk}$ is identical to $H_2O$. If the dissociation products of $H_2O$ are not considered a part of the $H_2O$ in a sample of $water_{nk}$, we've already shown that $water_{nk}$ contains more than $H_2O$.[36]

Someone might object to the third provision; but, a friend of [PWA-4] could argue that the hydrogen bonding that holds the individual molecules in a cluster gives water a self-organizing property. Identity theorists are free to deny this, of course; but, if the association products of $H_2O$ are not considered a part of the $H_2O$ in a sample of $water_{nk}$, we've once again shown that $water_{nk}$ contains more than $H_2O$.

Given these provisions as to what may be included in something aptly described as pure $H_2O$, [PWA-4] may be taken as true. But, then it follows that $water_{nk}$ can't be identical to $H_2O$ alone, [PWA-5]. If it were, then, by the transitivity of identity, it would be identical to purified water, contradicting [PWA-3].

Consequently, the conclusion is inescapable. We must reject a belief that is very popular among physicalist philosophers; namely, that $water_{nk}$ is identical to $H_2O$ *alone*.

Whether or not this conclusion turns out to be a fatal blow to the theory of the water/$H_2O$ relation, I consider it a victory for common sense. Is any philosopher tempted to say that iron ore is identical to the iron it contains? If not, why think that $water_{nk}$ is identical to the hydrogen oxide it contains? In each case, the original sample and the product of its purification have different properties. Consequently, they are distinguishable rather than identical.

---

36 A more technical reason for considering the hydroxide ($HO^-$) and hydronium ($H_3O^+$) ions an essential part of a sample of hydrogen oxide is that, if we held otherwise, we would not be able to explain all the facts about water by citing facts about $H_2O$. If one poured equal amounts of protium oxide and deuterium oxide into an empty container, the resulting mixture won't stay that way for long. Because hydrogen oxide molecules are constantly dissociating and randomly re-associating, the mixture quickly achieves an equilibrium which is about 25% protium oxide ($^1H_2O$), 25% deuterium oxide ($^2H_2O$) and 50% semi-heavy water ($^1H^2HO$) in which one hydrogen atom is protium and the other is deuterium.

* * *

Before turning to examine the constitution reading of [WH], I want to draw out an implication of the failure of the identity reading of [WH].

Few dispute that $H_2O$ is (nomologically) *necessary* for $water_{nk}$ – water as we know it here on actual Earth; but, if $water_{nk}$ is a mixture that includes more than what $H_2O$ includes, we must concede that $H_2O$ is *not sufficient* for $water_{nk}$. Hence, there can be no theoretical identity between the chemical kind, $H_2O$, and $water_{nk}$.

Nevertheless, we could say that $H_2O$ *is* sufficient for artificially purified water; so, on the assumption that the scientific definition of water defines something for which $H_2O$ is both necessary and sufficient, perhaps, we could say that there is a theoretical identity possible between $H_2O$ and artificially purified water. We may even be willing to stipulate that $water_{sk}$ is the scientific name for artificially purified water.

But, none of that would mean that $water_{sk}$ is identical to $water_{nk}$. Hence, the claimed theoretical identity between $water_{nk}$ and $H_2O$ is false.

The appearance of truth as to [WH-1] is an illusion; but, what is it due to? Offering to show a theoretical identity between a natural kind term, $water_{nk}$, and a scientific kind term, $H_2O$, but delivering only the identity between artificially purified water ($water_{sk}$) and $H_2O$ is just a bait and switch operation. Absent an equivocation as to the meaning of water it can not succeed.

### §3.3.3.6.4.4 The Theory of Constitution

According to some philosophers, the paradigmatic example of a necessary a posteriori identity

**[WH] Water is $H_2O$**

is ambiguous. It may be read as an identity claim or as a claim of constitution. Since the identity reading seems to lead us into absurdity, let us consider the constitution reading, its major alternative.

**[WH-2] Water is constituted by $H_2O$**

Despite having stated the defining claim of the constitution theorist, we can not yet proceed to examine the facts to see whether they support the theory. The problem is that we don't yet know what [WH-2] means.

By default the constitution reading is a theory of material constitution, which implicitly assumes that an object or substance is constituted by the material of which it is composed and nothing else. I will defend a minority view, the theory that a object or substance is constituted by its material *and* the form that material takes to be that object or substance. I call this the *theory of hylomorphic constitution*.

In what follows in this subsection, I will briefly review what advocates and reviewers have said about material constitution for the purpose of defining

where hylomorphic constitution differs from material constitution. In following sections, I'll review what advocates of theories of material constitution say about the diamonds/carbon relation and about the water/$H_2O$ relation with a view to showing that hylomorphic constitution is the better theory.

Clearly, a theory of material constitution and a theory of hylomorphic constitution differ in their assumptions as to the nature of an object/substance. However, there is less clarity as to the other commitments of the constitution view. Ryan Wasserman (2004, 693) writes that a constitution theorist is committed to three important claims:

> [CT-1] That ordinary objects (tables, chairs, lumps of clay and statues) exist.
>
> [CT-2] That ordinary objects have the modal properties attributed to them.
>
> [CT-3] That constitution is not identity.

[CT-1] rejects mereological nihilism, the position that only fundamental entities (eg. quarks and leptons) actually exist. What appear to be composite objects such as tables and chairs are merely collections of quarks and leptons arranged so that they appear to be tables and chairs. Contrary to the mereological nihilist, constitution theorists of both varieties, materialistic and hylomorphic, hold that lumps of clay exist and that they can be fashioned into statues.

Both varieties of constitution theorist would also say that the lump of clay would survive being squashed flat; whereas, the statue would not survive. However, it is not clear to me how our willingness to make such modal claims justifies believing that objects have modal properties.

*What are the grounds for denying that objects have modal properties?*

Consider an analogous situation. Does the use of color predicates by subjects entail the presence of color properties in objects? No. I might say that tomatoes are red; but, making such a claim does not entail that objects have color properties; so, predicate ascription does not entail property detection. I do not necessarily contradict myself by affirming that the tomato is red and denying color realism. I might argue that saying “that tomato is red” is merely a manner of speaking; and, that such a statement should be understood as meaning “that tomato looks red to me”, which does not suggest that the tomato has the property of being red. By analogy, I may state modal claims or express modal intuitions about objects; but, that doesn't entail that objects have modal properties.

As far as I know, scientists have never detected the modal properties of an object by empirical means; so, another basis for denying the existence of modal properties is simply that invoking properties detectable only by philosophers in the course of arguing for [CT-3] amounts to an assumption of property dualism, a controversial position which I reject.

In the case at hand, the clay and the statue, it's not necessary for me to deny the existence of modal properties because a theory of hylomorphic constitution has resources not available from a theory of material constitution.

Consider a typical example, the lump of clay and the statue into which it is fashioned. The clay survives being squashed flat. The statue does not. On the assumption that objects have modal properties, an advocate of material constitution may explain these facts by saying that the statue and the lump of clay have different modal properties. The constitution theorist invokes Leibniz to conclude that, “despite the fact that the statue and the lump stand in a very intimate relationship, they are nonetheless distinct” (Wasserman, 2004, 693); and, consequently, must explain the existence of distinct but materially and spatially coincident objects.

From the hylomorphic perspective, the lump of clay consists of the clay *and* the form that clay takes to be that lump. Materials such as clay can change their form. When a sculptor shapes the lump of clay into the form of, say, Hercules, we might want to say that sculptor has created a statue.

The statue came into being at the same time that clay took on the form required to be that statue. If the sculptor later decides to reform the clay into the likeness of Aphrodite, is there a problem? No. The statue of Hercules ceased to exist at the same time that the lump of clay ceased to have the form it was required to have to be the statue of Hercules.

The point is that, if we consider objects as compositions of matter and form, there is no problem with modal properties for the hylomorphic constitutionalist; so, I don't have to deny their existence until they can be used to generate an argument for non-identity of lump-in-form and statue.

There is also no problem of distinct but spatially and materially coincident objects for the hylomorphic constitutionalist. There is only one object; and, it is self-identical whether described as the statue of Hercules or a lump of bronze in the form of a legendary Hero. On the hylomorphic view, constitution preserves token identity.

Nevertheless, [CT-3] remains a point of considerable controversy.[37]

In the foregoing, I've assumed that we can read [CT-3] as if written “constitution is neither type nor token identity”. My impression is that constitution theorists take it this way; although, few say so as explicitly as Derk Pereboom.

> Mental natural kinds are not identical to natural kinds in physics because mental causal powers are not identical to microphysical causal powers. The fact that mental kinds are multiply realizable at the level of microphysical kinds provides an important clue as to why this is so. The nonreductivist view I propose departs from others insofar as it rejects the token identity of mental and microphysical entities of any kind – including causal powers. The deepest relation between the mental and the microphysical is material constitution, where this relation is not to be explicated by the notion of identity. (Pereboom, 2015, 428)

I will proceed by reviewing the attack on a restricted kind of type identity theory by an advocate of material constitution without identity. I will show that theory of hylomorphic constitution avoids the absurdities attributed to type identity

---

37 For general arguments that constitution is not identity, see Johnston (1992a) and Baker (1997; 1999). For arguments that constitution is identity, see Noonan (1993), Francescotti (2003) and Pickel (2010).

theories just as easily as does a theory of material constitution. I will also argue that hylomorphic constitution is compatible with some identity claims, making it the superior version of constitution theory.

### §3.3.3.6.4.5 Constitution – Identity or Non-Identity?

Mark Johnston puts the question nicely in his caricature of the Kripkean philosopher proposing marriage by offering his beloved "a nugget of gold and some chimney soot" (Johnston, 1997a, 564).

The philosopher explains that, according to Kripke, "each manifest kind of stuff is identical with the chemical kind which makes up the relevant samples of the manifest kind" (Johnston, 1997a, 564).

> … So, not only is the kind water numerically identical with $H_2O$, but so is the kind ice, the kind snow and the kind water vapor. Therefore by applying the transitivity of identity, we philosophers have discovered that snow is numerically identical with water vapor. ... The kind diamond is numerically identical to the chemical kind carbon, and so of course is soot. So diamond is identical with soot. (Johnston, 1997a, 564)

Clearly, there is something wrong with these tongue-in-cheek but very educational arguments that the philosopher presents to illustrate (what he takes to be) the Kripkean perspective. Let's consider the argument for the Identity of Diamond and Soot first, as it seems to be the simpler case.

[IDS-1] Diamond is identical to carbon

[IDS-2] Soot is identical to carbon

[IDS-3] (Therefore) Diamond is identical to soot

Clearly, [IDS-3] is absurd. However, if the premises are rewritten using the is of constitution the problem disappears.

[IDS-4] Diamond is constituted by carbon

[IDS-5] Soot is constituted by carbon

Nothing absurd can be deduced from [IDS-4] and [IDS-5]; consequently, Johnston argues that claims like "diamond is carbon" should be interpreted as claims of constitution rather than claims of identity.

In what follows, I will take Johnston's IDS argument through [IDS-3] as a decisive *reductio* of the specific claims, [IDS-1] and [IDS-2], that he is arguing against. However, refuting a single, cherry-picked type identity claim is not sufficient to justify the conclusion that no identity claim – not even a token identity claim – is possible.

*Is there an alternate identity claim not vulnerable to Johnston's objections?*

The identity theorist need only recall the point made in discussing the identity crisis of identity theory (§3.3.3.4.4). If we assume that a non-fundamental object (or substance) is constituted by matter *and* form, we can state the following (as a first approximation).

[Identity of Diamonds] A diamond is (identical to) a chunk of carbon arranged in the form carbon atoms must assume to constitute that

diamond.

If a materially constituted object is held to be constituted by its material alone, one may argue for the non-identity of an object and its constituents based on an evaluation of modal properties. Certainly, a diamond may be ground up and reduced to a collection of carbon atoms. The carbon atoms survive the destruction of the diamond; and, given [CT-2], that will count against the crude type identity claim of [IDS-1].

However, if a materially constituted object is held to be a composition of matter and form, the argument for non-identity based on an evaluation of modal properties will fail. When the diamond is reduced to a collection of carbon atoms, the diamond does not survive; but, neither does the form that those carbon atoms were in when constituting the diamond. Both the constitution claim and the identity claim stand or fail together.

It's difficult to see how a constitution theorist could object to claims such as [Identity of Diamonds]. Let's construct an analogous claim about soot.

> [Identity of Soot] A quantity of soot is (identical to) a quantity of carbon arranged in the form those carbon atoms must assume to constitute that quantity of soot.

Such identity claims do not lead to the absurdity of [IDS-3]; indeed, they allow us to derive its negation. Something that has the form carbon atoms must assume to constitute a diamond does not have the form carbon atoms must assume to constitute a quantity of soot. Hence, diamonds are not soot. If necessary, we could construct analogous arguments to demonstrate that a diamond is also not identical to a sheet of graphene or a quantity of buckyballs.

Now, although these are conclusions concerning non-identities, there is no basis for the conclusion that constitution is inherently a non-identity relation. One may plausibly argue that hylomorphic constitution is (a carefully chosen) identity, the self-identity of a thing (described as itself) and its constitution (the thing described as a composition of matter and form). Any given diamond is self-identical, whether I describe it as a diamond or as a mass of carbon atoms arranged in the crystalline structure those carbon atoms must assume to be that diamond.

At this point, the critic has a reply to the claim of constitution; but, it is not related to the question of whether the claim of constitution is treated as an identity or not. According to Johnston's own criterion of constitution a unique explainer is required.

> [JCC] If at an appropriately basic physical level carbon is the unique explainer of the causal powers characteristic of diamonds, then carbon constitutes diamonds. (1997a, 582)

Assuming that the causal powers of a diamond are its properties, there is a problem. If we want to attribute a color property to the diamond itself, we have to dispense with the assumption that carbon is the *unique* explainer of the properties of a diamond.

The color of a diamond doesn't come from the carbon it contains. A diamond

made of pure carbon is translucent with no hint of color, a white diamond. Some diamonds, however, have color. Diamonds can be yellow or blue or some other color. The color of a diamond is determined by the impurities that it contains and/or by deformations of the crystalline structure. A yellow diamond occurs when the crystalline structure is mostly carbon with a small amount of nitrogen. When the crystalline structure is mostly carbon with a small amount of boron, a blue diamond is the result. (Wikipedia, 2016-11-26).

By ignoring the crystalline structure in which the atoms composing a diamond are embedded, Johnston could argue that carbon is the unique explainer of the properties of white diamonds. Then, by [JCC], he could conclude that white diamonds are constituted by carbon alone; but, he would be unable to draw any conclusion about the constitution of colored diamonds from [JCC] alone.

In my view, we naturally want to say that a yellow diamond has a constitution of some sort despite not having a unique explainer of all of its properties. So, a theory of constitution capable of describing colored diamonds as well as white diamonds must dispense with [JCC]. According to the theory of hylomorphic constitution, an object or substance can be a mixture of materials; and, further, that a substance/object is constituted by all the materials that it contains *and* by the form in which they are arranged.[38]

Such a theory would allow us to say that a yellow diamond is constituted by the carbon and the nitrogen atoms it contains and the form in which those atoms are arranged, in this case the crystalline lattice necessary for those atoms to constitute that diamond. And something similar could be said about blue diamonds, the carbon and boron they contain and the form in which they are arranged.

It should be noted that a theory of constitution *must* include some mention of the arrangement of the materials said to constitute a material object. One reason for this was given above: to defeat claims for non-identity based on the difference between the modal properties of the constituted object and the constituting materials. There is another reason as well. Sometimes the arrangement of the atoms contributes to explaining ordinary (non-modal) properties of the substance or object.

> Pink and red [diamonds] are caused by plastic deformation of the crystal lattice from temperature and pressure. ... Purple diamonds are caused by a combination of crystal lattice distortion and high hydrogen content. (Wikipedia, 2016-11-26)

Now, we have a plausible theory of hylomorphic or material/form constitution. In

---

38 Such a theory of the material constitution of objects would be consistent with our expectations concerning substances, particularly liquids. Suppose I poured some vodka into some orange juice. The resulting mixture constitutes the mixed drink known as a screwdriver. While the ethanol in the vodka is the unique explainer of the intoxicating properties of a screwdriver, the orange juice is the unique explainer of its color. There is no one unique explainer of all the properties of the screwdriver; but, I'm not willing to say either that a mixed drink has no material constitution at all or that it is not constituted by the all the substances it contains. Once again, one naturally concludes that a unique explainer of properties is not required by a theory of constitution; and, further, that a substance/object includes whatever materials are required to explain its properties.

the case of diamonds, we may say that diamonds are constituted either by a crystalline structure of carbon atoms or by a crystalline structure of mostly carbon atoms with a small percentage of atoms of other elements. Now, however, a token identity claim certainly follows because a given diamond is self-identical whether it is described as a diamond or as a collection of carbon (and, possibly, other) atoms arranged in the crystalline lattice necessary for those atoms to constitute that diamond.

Having arrived at a closer approximation of what it means to be a yellow diamond, we can generate statements that seem to be about types.

> [D-1] A yellow diamond is a collection of mostly carbon atoms with some nitrogen atoms suitably arranged in a crystalline lattice.

It seems clear that [D-1] is a statement of type identity rather than type constitution because the terms in the subject and predicate positions can be reversed.

> [D-1.1] A collection of mostly carbon atoms with some nitrogen atoms suitably arranged in a crystalline lattice is a yellow diamond.

Constitution is generally said to be an asymmetric relation; so, if both [D-1] and [D-1.1] are true, that counts against interpreting the 'is' in either statement as the is of constitution.

One could recast *both* [D-1] and [D-1.1] as statements of class inclusion.

> [D-1.2] A collection of mostly carbon with some nitrogen atoms suitably arranged in a crystalline lattice is an instance of the type, *yellow diamond*.
>
> [D-1.3] A yellow diamond is an instance of the type, *collection of mostly carbon with some nitrogen atoms suitably arranged in a crystalline lattice*.

However, if both [D-1.2] and [D-1.3] are true that supports the conclusion that we can take [D-1] to be a statement about type identity. Let us also take the terms on each side of the 'is' in [D-1] to define sets of items; and, let us *hypothesize* that the two terms define co-extensive sets. If we tested this hypothesis empirically and found that the the two sets are in fact co-extensive, we would be justified in concluding that [D-1] was a true a posteriori identity.

*How would scientists test this hypothesis?*

I suspect that scientists would begin by constructing a suitable null hypothesis and then go look for evidence that it is false.

> [Null Hypothesis] The set of all yellow diamonds as identified by competent jewelers is ***not*** co-extensive with the set of all collections of mostly carbon atoms with some nitrogen atoms suitably arranged in a crystalline lattice as identified by competent scientists.

Now, let us imagine assembling a panel of expert jewelers and a panel of forensic scientists led by Abby Sciutto. Each is given items to examine and each is asked to record their decision as to whether the test item is or is not a yellow diamond. After all the test item examined, let us imagine that a statistical analysis of the test results indicated a very high correlation between the judgments of the

jewelers and the judgments of the forensic scientists. The resulting report would indicate that the null hypothesis was rejected with a high degree of statistical certainty.

We could then say that there was scientific evidence that yellow diamonds ***are*** collections of mostly carbon atoms with some nitrogen atoms suitably arranged in a crystalline lattice

Similar tests could be conducted for other kinds of diamonds. By combining the results of such tests, we could generalize from statements about kinds of diamonds to a statement about the kind, diamond.

> [D-2] A diamond is a collection of carbon atoms or of mostly carbon with some atoms of certain other elements suitably arranged in a crystalline lattice.

We could again imagine conducting an empirical test. Once again, I would imagine the null hypothesis being rejected with a high degree of statistical certainty.

* * *

*What has been accomplished?*

The rhetorical strategy of material constitution theorists seems to consist of showing that, given their assumption about the nature of an object – that it is constituted by its material alone – some seemingly plausible type identity claims (diamonds are carbon, soot is carbon) leads us into absurdity (diamonds are soot). But, the suggestion that these absurdities that can *only* be avoided by denying type identity in favor of type constitution is false. Such absurdities can also be avoided by less drastic means, simply by defining an object/substance as a constitution of matter and form.

The theory of hylomorphic constitutional for material objects/substances has the further merit of preserving token identity claims between Hercules and the lump/form of which it is constituted, thereby dispensing with the spatially co-incident objects that plague theories of material constitutional.

There is also the merit of preserving the common sense view. Suppose I disassembled your car in the dead of night. Even if I did not damage, hide or steal any of the parts, you would naturally feel that I had damaged your car because a car is more than a pile of car parts. It is a collection of car parts arranged so as to be a car.

Overall, hylomorphic constitution is the better theory of constitution.

*Where do we go from here?*

In the case of diamonds, anyway, it seems to be possible to work up from token identity claims to claims about the identity of types of diamonds; so, that may be yet another merit of hylomorphic constitution. However, there is no guarantee that similar type identity statements can be constructed in the case of water/$H_2O$.

### §3.3.3.6.4.6 The Constitutional Readings of Water/$H_2O$

To review, the paradigmatic example of a necessary a posteriori identity.

**[WH] Water is $H_2O$**

is ambiguous. It may be read as an identity claim or as a claim of constitution.

**[WH-1] Water is identical to $H_2O$**

**[WH-2] Water is constituted by $H_2O$**

I've presented reasons for thinking that [WH-1] is false; and, Johnston would certainly agree with this. He argues that we should “understand the relation between manifest kinds like water and chemical kinds like $H_2O$ to be constitution rather than identity”. (Johnston, 1997a, 566)

After developing a theory of hylomorphic constitution I revisited Johnston's (tongue-in-cheek) argument for the Identity of Diamonds and Soot and showed that the absurdities Johnston identified can be banished with less collateral damage. I will now consider the following simplified version of Johnston's argument for the Identity of Water and Ice.

[IWI-1] Liquid water is $H_2O$

[IWI-2] Ice is $H_2O$

[IWI-3] (Therefore) Liquid water is ice

Now, Johnston would point out that we can avoid the absurdity of [IWI-3] as easily as we avoided the absurdity of [IDS-3] (considered in a previous section), simply by reading [IWI-1] and [IWI-2] as claims of constitution.

[IWI-4] Liquid water is constituted by $H_2O$

[IWI-5] Ice is constituted by $H_2O$

As with the diamond/carbon example, nothing absurd follows from [IWI-4] and [IWI-5]; so, once again it looks like the theory of material constitution without identity helps us avoid the absurdities that follow from claims of crude identities; but, a theory of hylomorphic constitution will also have the pragmatic effect of avoiding those same absurdities.

By adopting a theory of hylomorphic constitution in which an object or substance is constituted by matter and form, we aren't troubled by the absurdities that follow from assuming that an object or substance is identical to its material constituents alone. As in the case of diamonds/carbon, we may easily recover a token identity claim.

Any given chunk of ice is self-identical whether it is described as ice or as a sample of $H_2O$ molecules that have assumed the crystalline lattice form they must assume to constitute or be ice. And, similarly for liquid water and water vapor.

*How do we decide between material constitution and hylomorphic constitution?*

One way would be to challenge the essentialist assumption that Johnston built into his description of the Kripkean position. Here it is again, this time with my

emphasis on its vulnerability: “each manifest kind of stuff is identical with ***the*** chemical kind which makes up the relevant samples of the manifest kind” (Johnston, 1997a, 564 (emphasis supplied).

Johnston preserves the single chemical theory of water in his constitutional reading as shown by the importance he attributes to having a unique explainer in defining constitution. His principle, (27*), which I've renumbered is:

> [MJ-27] If at an appropriately basic physical level $H_2O$ is the unique explainer of the causal powers of water, then $H_2O$ constitutes water.

Consequently, it appears that Johnston is interpreting [WH-2] as if it read something like ...

> **[WH-2.1] Water is constituted by $H_2O$ *alone***
>
> **[WH-2.2] Water is constituted by $H_2O$ *and nothing else***
>
> **[WH-2.3] Water is constituted by *nothing other than* $H_2O$**

Or perhaps, since he spoke about samples of water, he is interpreting [WH-2] as something more like a token identity claim, something like ...

> **[WH-2.1.1] A sample of water is constituted by its $H_2O$ *alone***
>
> **[WH-2.2.1] A sample of water is constituted by its $H_2O$ *and nothing else***
>
> **[WH-2.3.1] A sample of water is constituted by *nothing other than* its $H_2O$**

[WH-2.2] and [WH-2.2.1] each have the virtue of capturing the view of U.T. Place, who called his theory an identity theory but stated it in terms of the 'is' of composition. In explaining what he meant by that, he distinguished the 'is' of composition from the 'is' of predication by saying that with the former but not the latter, it was possible to add the phrase “and nothing else” without changing the meaning (1956, 45). [WH-2.3] and [WH-2.3.1] each have the virtue of capturing the view of Frege.

Johnston holds that [MJ-27] is “plausibly taken as priori” because it follows from what we should mean by "material constitution” (Johnston, 1997a, 582). The problem is that [MJ-27] is arguably false. Consider a reading of [MJ-27] that is explicitly about $water_{nk}$.

> [MJ-27.1] If at an appropriately basic physical level $H_2O$ is the unique explainer of the causal powers of $water_{nk}$, then $H_2O$ constitutes $water_{nk}$.

The antecedent is false if “water” is taken to mean $water_{nk}$. $H_2O$ is not the unique explainer of the properties of $water_{nk}$. We need to consider the minerals dissolved in $water_{nk}$ to explain the degree to which it conducts electricity. Further, the consequent is also false on such a reading. A sample that is constituted by $H_2O$ and nothing else would be purified water not $water_{nk}$.

The material constitutionalist may reply that the question turns on the definition of constitution. Setting aside the question of whether the form or arrangement of the material is also important, I am defining constitution so that an object or a

sample of a substance is constituted by all the material(s) it contains. The material constitutionalist may say that a substance is constituted only by its dominant constituent.

It would seem arbitrary and *ad hoc* to adopt such a position only with respect to water; but, could it plausibly be applied to other materials which are mixtures? Brass is an alloy consisting of 2/3 copper and 1/3 zinc. That makes copper the dominant constituent. Should we conclude that copper is the unique explainer of the properties of brass, that brass is constituted by its copper alone or that brass is copper? I don't think so.

Suppose we revisit the case of the sculptor who casts a statue out of bronze consisting of 88% copper and 12% tin. Material constitutionalists would like to say that the statue is constituted by the bronze. Would they say that the statue is constituent only by the copper that constitutes the bronze? I don't think so. As far as I know, no philosopher has (yet) taken that position in the literature on statues and the materials out of which they are made.

So, why think that $water_{nk}$ is constituted only by its predominant ingredient?

In short, all of the constitutional interpretations of [WH] listed above fail for the same reason the corresponding identity readings of fail: water is a mixture.

As an alternative, I propose the following:

> **[WH-2.4] A sample of water is constituted by its material, either $H_2O$ alone or $H_2O$ plus other substances that give the resulting mixture the properties that we attribute to that sample, and the form in which its material(s) are arranged to be that sample.**

### §3.3.3.6.4.7 Consequences: The Explanatory Gap Argument

In the traditional analysis of the explanatory gap there is an asymmetry between the (alleged) identities, water/$H_2O$ and pain/C-fibers. Facts about $H_2O$ explain or at least contribute to explaining facts about water but facts about firing C-fibers don't explain how or why an experiential phenomenon is associated with those firing C-fibers. Thus, while one might plausibly deny the identity of pain and firing C-fibers, one could not plausibly deny the identity of water and $H_2O$.

Johnston argues that we must deny the identity of water and $H_2O$ to avoid the absurdity of holding that liquid water is identical to ice; and, that a theory of material constitution without identity is the only alternative. If both water/$H_2O$ and pain/C-fibers are cases of material constitution, as Johnston assumes, there is no asymmetry between the two cases and no explanatory gap to be explained or explained away, or so Johnston argues.

One might follow Johnston in denying the identity of water and $H_2O$; but, disagree as to whether this conclusion resolves the debate over the explanatory gap. Benbaji (2008), for example, argues that even assuming that both cases are case of material constitution, explanatory gap arguments are not undermined.

In one sense, I have pursued a more decisive reply, preserving the asymmetry

between the cases of water/$H_2O$ and pain/C-fibers. In contrast to Johnston's theory of material constitution, a theory of hylomorphic constitution allows us to state a token identity claim in the case of a sample of water and that which it contains; but, that no such token identity claim is possible in the case of pain and C-fibers because they have different modes of existence.

This points to a further vulnerability in Johnston's account, the assumption that pain has a material constitution.

Some theory of constitution will work for the water/$H_2O$ relation because no one disputes that water is a material substance or that $H_2O$ is a material substance. The only real questions are which theory of constitution provides the best account of the water/$H_2O$ relation and whether it is constitution with or without identity.

In the case of pain/C-fibers, Johnston makes no attempt to justify assuming that that pain is a material object or a material substance; so, assuming that pain has a material constitution is obviously question begging.

### §3.3.3.6.4.8 A Better Paradigm of an A Posteriori Identity?

Water is an extremely complex substance; perhaps, too complex for a claim as simplistic as [WH-1] to be an illuminating example of a necessary a posteriori identity. But, if we reject the identity of $water_{nk}$ and hydrogen oxide, $H_2O$, would identity theorists be able to provide a better example of an a posteriori identity?

There are certainly other examples of alleged identities that we learn about from experience and empirical research; but, arguments for each proposed identity would have to be examined for assumptions beyond the empirical facts; for example, the assumption introduced by Beckermann, that $water_{nk}$ is a single chemical kind.

*Are such assumptions necessary to reach a conclusion about $water_{nk}$ and $H_2O$?*

In my view, yes. Beckermann introduced his essentialist assumptions to reach the conclusion that water is identical to $H_2O$. Johnston introduced his essentialist assumptions to reach the conclusion that constitution is not identity. I introduced the assumption that a material object or substance is matter-in-a-form, matter arranged a certain way, in order to reach the conclusion that some identity claims are compatible with a theory of hylomorphic constitution.

One may reasonably wonder whether introducing such assumptions undermines the claim of identity and/or the claim of necessity.

#### §3.3.3.6.4.8.1 Introducing Essentialist Assumptions

Introducing essentialist assumptions has much the same effect as introducing reductive assumptions. The concept being 'essentialized' is transformed just as the concept being reduced is transformed before the statement of identity or constitution is derived.

This can be seen by considering the structure of arguments for alleged reductions. The first stage consists of modifying the concept of that which is to

be reduced so that the process may proceed. Levine puts it this way

> Stage 1 involves the (relatively? quasi?) a priori process of working the concept of the property to be reduced 'into shape' for reduction by identifying the causal role for which we are seeking the underlying mechanisms (Levine, 1993, 132).

Kim is a bit more circumspect. He writes that someone who wants to reduce an allegedly emergent property E to some property in the domain of its reduction base, B, must first functionalize E.

> ... that is, E must be construed, or reconstrued, as a property defined by its causal/nomic relations to other properties, specifically properties in the reduction base B. (Kim, 1999, 10)

This model of reductive analysis seems to be agnostic as to whether the reduced property is identical to or is constituted by the reducing property.

> An important part of this procedure is to decide how much of what we know (or believe) about E's nomic/causal involvement should be taken as defining, or constitutive of, E and how much will be left out. (Kim, 1999, 11)

The next step involves "Find a theory (at the level of B) that explains how realizers of E perform the causal task that is constitutive of E". (Kim, 1999, 11)

* * *

Introducing essentialist assumptions results in a similar transformation of the concept being reduced or essentialized. But, if making the argument work requires introducing essentialist or reductive assumptions, it undermines the claim that [WH] an a posteriori *identity*.

As was discussed in the analysis of the Purified Water Argument, the identity claim that a posteriori physicalists claim to have shown is the theoretical identity between a natural kind, $water_{nk}$, and a scientific kind, $H_2O$. However, if $water_{nk}$ must first be essentialized or reduced to artificially purified water before an identity is shown between artificially purified water and $H_2O$, the promised theoretical identity has not been demonstrated.

What's worse is that the *non-identity* of $water_{nk}$ and artificially purified water precludes the identity of $water_{nk}$ and $H_2O$.

Worse, the Kripkean approach holds that [WH-1] is an a posteriori *necessity*; meaning, that it is inconceivable that it should be false in any possible world. However, it is more than merely conceivable that someone might deny the essentialist and/or reductive assumptions that are build into an argument for an a posteriori identity. Indeed, it's more than merely possible.

It's actual.

I deny any assumptions that allow essentializing or reducing $water_{nk}$ to purified water in the course of "demonstrating" that $water_{nk}$ is identical to $H_2O$ alone.

### §3.3.3.6.4.8.2 A Lesser Necessity?

Since it is not necessary for anyone to add reductive assumptions to the

empirical facts, the claim that "water$_{nk}$ is identical to $H_2O$ alone" is a *necessary* a posteriori identity becomes especially dubious. One may deny any of those additional assumptions without perpetrating a logical contradiction; so, how could a conclusion resting on those assumptions be *metaphysically* necessary?

In my view, we are not dealing with metaphysical necessity but with a lesser degree of necessity that I will call *analytic necessity*.

Assume that there are two adult human philosophers, Jack and Jill, male and female, respectively, who are each unmarried. Given the definition of "bachelor" as an unmarried male, it is analytically true that Jack is a bachelor and that Jill is not. However, we can easily imagine a possible world which is a near duplicate of our actual world but in which "bachelor" has been defined to refer to a rare breed of sea turtle. In such a world, neither Jack nor Jill would be a bachelor. Consequently, the truth that Jack is a bachelor is not metaphysically necessary. It is only analytically necessary that, in our world (that part of it that speaks English, anyway), Jack is a bachelor.

Scientists sometimes pause their investigations to define their terms to reflect what they already know. An interesting example of this process occurred in 2006 when the International Astronomical Union's General Assembly approved a resolution, Resolution 5A, classifying the celestial bodies other than satellites orbiting the Sun into three categories: planets, dwarf planets and small solar system bodies (Wikipedia, 2017-01-15).

According to the criteria for classifying objects, Pluto is a dwarf planet.

Now, the term *dwarf planet* itself suggests that a dwarf planet is a planet must as a dwarf star is a star, a dwarf galaxy is a galaxy and a dwarf human is a human. There was even a secondary resolution, 5B, introduced to specify that a dwarf planet is a subtype of planet; but, this resolution was defeated. While the IAU never officially approved a third resolution stating that a dwarf planet is not a subtype of planet, the defeat of Resolution 5B could be taken to mean that a dwarf planet is not a planet; so, let us take it that way.

Now, given the IAU definitions and the empirical facts, it is true that Pluto is a dwarf planet but not a planet; but, this is only an analytic truth, something true because of the way we've defined our terms. This is not a metaphysical truth because it is quite easy to imagine a possible world in which Resolution 5B was approved making dwarf planets a subtype of planet.

Who knows what will happen on actual Earth in the near future. Astronomers may approve a resolution making dwarf planet a subtype of planet. Then the same empirical facts plus the new definition will make Pluto a planet again. It will again be an analytic truth, necessary after a fashion; but, it will not be a metaphysical necessity.

Similarly, we can imagine a situation in which the chemists of the world convene and confer and then issue a press release announcing that they've defined "water" to be a single chemical kind, $H_2O$. It would now be necessarily true that water so defined is identical to $H_2O$ alone. However, it would only have the necessity that comes from an analytic truth, something true only because of the way we've defined our terms.

The role of analytic truth in some classic examples of a posteriori identities is greatly underappreciated. Consider [V-3] which is generally taken to be a metaphysically necessary a posteriori identity.

### [V-3] The Morning Star is the Evening Star

Given definitions of “The Morning Star” and “The Evening Star” making those terms names for the celestial object also named Venus, we would conclude that [V-3] is true. Now, in the usual Kripkean analysis, [V-3] is true in all possible worlds in which the terms refer because the terms are rigid designators of the object to which they allegedly refer.

However, it is only by definition that the terms *Morning Star* and *Evening Star* refer to an object rather than to an appearance of an object.

I have adopted different definitions according to which *The Morning Star* is the name of the appearance to humans on Earth of Venus in the morning sky and *The Evening Star* is the name of the appearance to humans on Earth of Venus in the evening sky.[39]

Now, [V-3] is false because the appearances named are distinct.

Nevertheless, the a posteriori fact discovered by astronomers is still respectable. I simply say that the object of which The Morning Star is an appearance is the same object as the object of which the Evening Star is an appearance; and, that object is necessarily self-identical.

The self-identity of an object may be a metaphysically necessary truth; but, the claim that [V-3] is true depends on both the empirical facts and a linguistic convention defining the terms used in that statement. With a different linguistic convention, [V-3] would be false, the empirical fact would still be true (although it would be expressed differently) and the law of identity would still be a priori.

The claim that [V-3] is true and the claim that [V-3] is false turn on the definition of the terms used; so, the most we can say is that [V-3] is analytically true or false, depending on how we define our terms.

This is a seriously deflationary view of the so-called necessary a posteriori identity. For many philosophers, the brain/experience identity that they allege is a metaphysical necessity rather than an analytic necessity. Anything less than metaphysical necessity, nomological necessity, for example, concedes something to zombie arguments that identity theorists probably don't want to concede. If the only necessity was analytic necessity, even more would be conceded.

Consequently, a careful evaluation of the degree to which alleged examples of necessary a posteriori identities depend on the way we define our terms reduces the likelihood that a posteriori identities will be of much use in defending identity theory physicalism. Nevertheless, there is at least one identity theorist who admits *aiming* for an analytic necessity, Ullin T Place.

---

39 These definitions make the case of the Morning/Evening Star analogous to the case of the Full/Crescent Moon. The Moon is the Moon; but, the Full Moon is not identical to the Crescent Moon because they are distinct appearances. Similarly, Venus is Venus; but the Morning Star is not identical to the Evening Star because they are distinct appearances.

> But whereas the typical token-identity statement, “His table is an old packing case,” if true, is contingent and synthetic, the typical type-identity statement of which “Water is $H_2O$” is a paradigm case is necessary and analytic. (Place, 1999, 82)

Recognizing that analytic input is as important as a posteriori input in generating the type of statements that are commonly known as a posteriori identity statement may open a can of worms some philosophers would prefer not to open: the possibility that the knowledge involved in these cases is analytic a posteriori knowledge.,

Stephen Palmquist has done more than anyone else to make the concept of analytic a posteriori knowledge respectable.[40]

> Analytic knowledge is that knowledge whose propositional expression is true (given an accepted meaning for the terms involved) solely by virtue of logical laws. … analytic a posteriori knowledge would be conceptual and contingent, if it were possible. (Palmquist, 1987, 252)

In the case of [V-3], the knowledge that The Morning Star is identical to the Evening Star is conceptual in that it relies on definitions of the terms (the names) involved but contingent in that it is only a contingent fact about our world that “The Morning Star” and “The Evening Star” are generally defined to be names for objects rather than names for appearances. However, it is not hard to imagine a possible world in which these terms are defined as names for appearances. In such a world, [V-3] would be considered false; although, of course, it would be granted that the object of which The Morning Star is an appearance is the same object as the object of which The Evening Star is an appearance.

The way the outcome – whether [V-3] is true or false in the actual or some possible world – turns on how we define a name. For Palmquist, naming is a good indicator of the analytic a posteriori.

> Accepting the analytic a posteriori as a legitimate epistemological category enables us to distinguish, in a way which neither Kant nor Kripke succeeded in doing, between the status of "naming" and "defining". To name requires that we adopt a practical perspective, according to which we act "as if" (or stipulate that) a certain object is to be rigidly designated by a certain word. That is, we subsume an object as experienced (a posteriori) under a given concept (analytically). To define, by contrast, requires that we adopt a logical perspective, according to which we devote all our attention to accumulating a set of properties which describe a concept uniquely. (Palmquist, 1987, 271)

The currently most popular naming convention guarantees that [V-3] will appear true to all those who stipulate that “The Morning Star” and “The Evening Star” refers to the object we also know as Venus. But it is surely a contingent fact about a world that one naming convention or another is more popular among the analytic philosophers of that world.

Palmquists takes a more general approach when he contrasts the analytic a posteriori with the synthetic a priori.

---

40

> The synthetic a priori and analytic a posteriori are therefore similar classes of knowledge insofar as both are concerned with conditions imposed on the world by the subject ... but they differ by virtue of the fact that one imposes *general* conditions (a priori) with intuitive (synthetic) content, whereas the other imposes *particular* conditions (a posteriori) with conceptual (analytic) content. (Palmquist, 1987, 273)

This suggests that the choice between a theory of material constitution and a theory of hylomorphic constitution is a choice between alternate analytic a posteriori truths. In explaining why diamonds are not soot or why lumps of clay can be fashioned into statues and then squashed flat, we've imposed a definition of an object on our understanding of all these thought experiments.

According the the theory of material constitution, an object is the material of which it is composed; whereas, according to the theory of hylomorphic constitution, an object is the material of which it is composed in the form that material must assume to be that object. Philosophers holding these theories may reach different conclusions (as, for example, Johnston and I reached different conclusions).

This leads us to one conclusion about analytic a posteriori truths. Since they follow from definitions one has *chosen*, they are contingent facts about a given philosopher. Which definitions are more popular in our actual world or in some other possible world is a contingent fact about that world.

A final example of what may be taken as analytic a posteriori knowledge concerns the classification of items such as phenomenal color, auditory and, taste sensations and so on. Our knowledge of experience is a posteriori if anything is. Intuiting or judging that aspects (bits and pieces) of experience falls under one concept (e.g. phenomenon) rather than another concept (e.g. property) is imposing a concept on experience; hence, our knowledge that experience consists of experiential phenomena or phenomenal properties, as the case may be, would qualify as analytic a posteriori knowledge.

I have argued for classifying aspects of experience as phenomena. However, I do not have a deductive argument beginning with uncontroversial first principles and concluding that aspects of experience are phenomena and not properties; and, I'm not aware of any analogous argument purporting to show that aspects of experience are properties and not phenomena. Philosophers may assume one classification or the other as a matter of choice. We could say that they have adopted competing philosophical intuitions; or, less charitably, that they have pragmatic reasons for their choices – as we've seen, if the outcome of the debate turns on the appearance/reality distinction, the language of discourse affects the outcome of the debate.

What we end up with is competing theories that each appear true to those who hold the required linguistic conventions. This perspective allows us to recognize that we are dealing with conflicting relative truths. These conflicting truths emerge when choices are made.

We've seen indicators that may help us identify the point where the philosopher must make such a choice. For example, suppose that scientists and philosophers build the case for parallel phenomenology to the point where Clark (1994) is

willing to *argue* for an inference to what he considers the best explanation for parallel phenomenology, identity. Unless this leap of choice is compelled by the empirical evidence, one may simple choose an alternate explanation; for example, that information that is physically instantiated in the brain may become phenomenally instantiated in experience.

As Feigl notes, this step from parallelism to identity is a matter of interpretation; or, I suggest, a matter of intuition and choice.

Should a choice such as this become very popular we may find ourselves in the situation Place described so well in the case of the brain/consciousness relation.

> … whereas token identities are typically synthetic, contingently true, if they are true and verified empirically, type identities are typically analytic, necessarily true in the sense that their denial is self-contradictory and true a priori. The focus of interest is on the conditions under which type-identity statements are synthetic, contingent, and subject to empirical verification, namely, the conditions that obtain in the case of the proposed type identity between consciousness and some as yet unspecified brain process. In such cases the fact that the two predicates invariably have the same extension remains to be demonstrated. Only when it is, and the ensuing identity statement becomes a matter of common knowledge, will it become an analytic, necessary, and a priori truth. (Place, 1999, 87)

### §3.3.3.6.4.9 Consequences for Knowledge Arguments

A constitution based theory of the relation between experiential and physical phenomena could easily appear to be reductive; provided, that one loosens the criteria for reduction so that an identity is not required. This may suffice for some physicalists and for some purposes; particularly, where the purpose of making a constitutional claim is to avoid allegations of dualism that might seem to follow from an admission of non-identity. However, in the context of the KA, affirming a constitutive relation between experiential and physical phenomena effectively waives some defenses identity theorists might have.

If the instance of phenomenal redness that Mary experiences seeing when she leaves her room and gazes upon a ripe tomato is not identical to NFP-4738 but only constituted by it, there is no possibility of arguing that Mary merely discovered an old fact appearing to her in a new guise.

Relying on a definition to reach a conclusion raises interesting questions and will likely have its own consequences. For one thing, the status of the conclusion as a posteriori knowledge is in peril. It is rather unlikely that empirical research will compel adoption of one definition of “material object” instead of some other. Secondly, a conclusion that rests on a definition is true analytically if true at all.

### §3.3.3.6.5 Papineau and the Causal Argument for Materialism

As noted earlier, Papineau recognizes an intuition of distinctness which impels us to view the physical and the phenomenal as distinct. Unlike Almog, Papineau doesn't *explicitly* recognize a competing intuition; although, he clearly recognizes that we are pulled in two directions.

> On the one hand, it seems clear that consciousness must be a normal part of the material world. Conscious states clearly affect our bodily movements. But surely anything that so produces material effects must itself be a material state.
>
> On the other hand, it seems absurd to identify conscious states with material states. Conscious states involve awareness, feelings, the subjectivity of experience. How could mere matter on its own account for the miracle of subjective feelings? (2002, 1)

The second *pull* is clearly the intuition of distinctness; and, Papineau is candid about its pervasive influence. It promotes dualistic thinking among those seduced by it; and, it moves materialists to admit that, even if it materialism was known to be true, it would be difficult to believe.

The first pull is not identified; but, it seems that something like Almog's intuition of materiality motivates the stance Papineau takes concerning the causal closure – he calls it the "completeness" – of physics. Without causal closure "there is no compelling reason to identify the mind with the brain". (2002, 255)

Papineau introduces several versions of his Causal Argument for Materialism, CAM, in the course of his book; but, to illustrate the role played by the causal closure principle, we only need to consider the first version (2002, 17-18).

> **[CAM-1] Conscious mental occurrences have physical effects.**
>
> **[CAM-2] All physical effects are fully caused by purely physical prior histories.**
>
> **[CAM-3] The physical effects of conscious causes aren't always overdetermined by distinct causes.**

[CAM-1] is unobjectionable.

[CAM-2] prohibits causal underdetermination, the situation in which an effect is the joint effect of a mental cause and a physical cause where neither by itself would be sufficient to produce the effect. I object to [CAM-2] on the grounds that it denies the most plausible model of mental causation.

[CAM-3] prohibits systematic overdetermination, the situation in which an effect is the joint effect of a mental cause and a physical cause where each alone is sufficient to produce the effect. I do not assume systematic overdetermination; so, I have no objection to [CAM-3].

Papineau says that the most obvious examples of [CAM-1] "are cases where our conscious feelings and other mental states cause our behavior" (Papineau, 2002, 17). Later, he gives a couple of specific examples while explaining that epiphenomenalism is unattractive because

> … it would require us to deny many apparently obvious truths, such as that my conscious thirst caused me to fetch a beer, or that my conscious headache caused me to swallow an aspirin. (Papineau, 2002, 22)

Presumably, my conscious headache is a headache of which I am conscious (rather than a headache which is itself conscious). If so, I deny Papineau's obvious truth that my headache *caused me* to swallow an aspirin; but, I'm not saying that my headache is epiphenomenal. In my view, an experiential phenomenon has a causal role; but, its role need not be very elaborate. All it

really needs to do is bring itself to my attention so that I may exercise *my* causal role, making choices and taking action based on reasons.

Upon noticing my headache, I briefly considered taking a nap, the best remedy for a headache, in my opinion. But I decided to take something for it. I then considered what I might take. I keep aspirin and acetaminophen in stock at all times; and, sometimes ibuprofen as well. This time I decide to take a combination of aspirin and acetaminophen.

Attributing to the headache the power to cause me to take an aspirin seems to move us in the direction of determinism (for headaches and afterimages) and epiphenomenalism for consciousness qua subject of experience. Moreover, it denies the phenomenology of willing. It seems to me that I make choices once I become aware of a need to take action; consequently, I deny that headaches have the power to determine outcomes – to cause me to take an aspirin.

However, according to [CAM-2], choices that are uncaused by a prior physical cause can have no physical effect unless they are themselves physical causes; hence, any mental cause must be identical to some physical cause.

What is the argument for [CAM-2]?

Ultimately, the principle of the causal closure of physics rests on the dubious allegation that this principle is supported by scientific evidence.

Papineau provides an extensive historical survey of scientific and philosophical thinking as to the causal completeness of physics. Papineau's aim is to show that physicalism is a reflection of developments in empirical theory rather than an intellectual fad.

> But once the completeness of physics became part of established science, scientifically informed philosophers realized that this crucial premiss could be slotted into a number of variant arguments for physicalism. There seems no reason to look any further to explain the widespread philosophical acceptance of physicalism since the 1950s. (Papineau, 2002, 255-6)

Clearly, this account of the rise of physicalism supposes that philosophers noticed the developments in physics; adopted the causal closure principle to maintain consistency with empirical results; and, accepted the identity of mind and brain as the logical consequence of the empirical evidence.

Incredibly, this is supposed to have taken place (among scientifically informed philosophers!) after developments in quantum mechanics resulted in

1. The collapse of the Laplacian determinism that seemed to follow from classical physics; and,
2. The recognition of an astounding possibility, that physicists have had to postulate consciousness to explain the laws of physics. (“It was not possible to formulate the laws of quantum mechanics in a fully consistent way without reference to the consciousness.” (Wigner, 1962, 169))

Arguably, the causal closure principle became popular *despite* the trend of empirical research. In any case, Papineau seems to acknowledges the fragility of his argument. In the final footnote in Thinking about Consciousness he wrote:

> On some, but not all collapse interpretations, distinctive special causes will be responsible for whether a collapse occurs or not (even though the subsequent chances of the various possible outcomes will still depend entirely on prior physical forces). I am thinking here of interpretations which say that collapses occur when physical systems interact with consciousness (or indeed which say that collapses occur when there are 'measurements' or 'macroscopic interactions', and then refuse to offer any physical explanations of these terms). On these interpretations, the completeness of physics will be violated, as well as the conservation of energy, since collapses don't follow from more basic physical laws, but depend on 'emergent' causes. It would seem an odd victory for anti-materialists, however, if the sole locus of sui generis mental action were quantum wave collapses. (Papineau, 2002, 255)

In the years since Thinking about Consciousness was written, developments in physics have further undermined support for the causal closure principle. The Conway/Kochen Free Will Theorem (2006, 2009) and similar theorems (Colbeck & Renner, 2011, 2012, 2013a, 2013b; Pusey, Barrett & Rudolph, 2012) *assume* that the physicist has the freedom to choose which experiment to perform and how to arrange the experimental apparatus.

Whether for this reason or some other, Papineau seems to have shifted his position. He now thinks that causation is an emergent phenomenon not present at the level of fundamental physics. At the most fundamental level, where all there is are particles in motion, physical laws are time symmetric. We have to go up a level of description, throwing away some detail in the process, to find causal effects (which are not time symmetric).

For this reason, "Mental properties can't be identified with strictly neural properties". Excluding the mental from being causes is wrongheaded because mental states are causes in their own right – independent of their neural realizers. "It's the mental that are causes and the neural events aren't." (2015a, @41:30).

Papineau sums up his new position, thus

> Once you realize that causation is an emergent phenomenon relative to basic physics, then we can understand how there can be autonomous mental causation in a material world. (2012, @46:00)

In another recent presentation, this one for the 2015 conference in Helsinki, *Toward a Science of Consciousness*, he explicitly addressed the implications of this new perspective for the causal closure principle.

> The standard argument for physicalism assumes the causal closure of the physical, based on the absence of ‘special’ forces in fundamental physics. But there are good reasons to suppose causation is an emergent macroscopic phenomenon, in the thermodynamic sense, akin to the increase of entropy. If this is right, then facts about fundamental physics have no direct implications for macroscopic causal patterns. And indeed there are independent reasons to think that that the physical realm is not in fact causally closed. (Papineau, 2015b)

With this Papineau has abandoned the principle that he once thought was the only basis for claims of physical/phenomenal identity.

### §3.3.3.6.6 Jackson's Identity Theory

Functional arguments for the identity of mental states and physical states revolve around the idea of the causal role. Jackson (2012) presents the general scheme for such arguments

> [CRA-1] Mental state M = The state that plays causal role C.
>
> [CRA-2] The state that plays causal role C = Brain state B.
>
> [CRA-3] Mental state M = Brain state B

[CRA-3] follows from [CRA-1] and [CRA-2] by the transitivity of identity; so, the argument appears to be valid. However, its soundness is questionable for reasons similar to those offered against other arguments involving talk of states. It is not clear how “state”, “mental state” and “brain state” are being defined.

We can differentiate physical states of the brain by the physical phenomenon that makes a brain state the physical state that it is; so, for example, a brain that is exhibiting or undergoing neural firing pattern, NFP-4738, would be in the physical state of exhibiting/undergoing that physical phenomenon.

Now, we know from Mary's Concordance of Phenomenology and Terminology that the experiential phenomenon associated with NFP-4738 is *tomato red*, the name of a certain shade of phenomenal redness. What we need to know is whether my brain exhibits (undergoes or instantiates) *tomato red*; but, all we have for evidence is that my brain exhibits (undergoes or instantiates) NFP-4738 while I am experiencing *tomato red*.

We can't just assume that tomato red is a state my brain is in or that tomato red is an experiential phenomenon exhibited by my brain. The state in which I am experiencing *tomato red* strikes me as a state I am in. It does not strike me as a state my brain is in. Of course, it may be that the state I am in is *also* a state my brain is in; but, in an argument for an identity claim, one can't just assume the identity one is trying to prove.

It gets worse. Unless the state I am in by virtue of experiencing *tomato red* is identical to the state my brain is in by virtue of exhibiting NFP-4738, it is not clear which state is the mental state referred to on the left side of [CRA-1].

I am willing to assume that the state I am in by virtue of experiencing tomato red is a mental state; and, while it's possible that Jackson would be willing to do the same, I'm not sure of that. In any case, absent an argument to the effect that the state I am in by virtue of experiencing *tomato red* is identical to the state my brain is in by virtue of exhibiting NFP-4738, I'm not willing to assume that the state my brain is in by virtue of exhibiting NFP-4738 is a mental state.

So, at this point, there are at least two variant readings of [P-1].

> [CRA-1.1] The state my brain is in while instantiating physical phenomenon P = The state that plays causal role C.
>
> [CRA-1.2] The state I am in while instantiating experiential phenomenon E = The state that plays causal role C.

Whether [CRA-1.1] or [CRA-1.2] is the canonical meaning of [CRA-1] would be a

highly controversial topic; they do not work equally well with [CRA-2].

Tweaking [CRA-2] a bit to make explicit the role of empirical research yields another pair of readings

> [CRA-2.1] The state that plays causal role C = the state my brain is in when it is performing its causal role, brain state B.
>
> [CRA-2.2] The state that plays causal role C = the state I am in when I am performing my causal role, brain state B.

Clearly [CRA-1.1] works well with [CRA-2.1]. It seems plausible to theorize that empirical research will eventually show that the brain performing its causal role is the brain in the state of instantiating some physical phenomenon, P. However, [CRA-3] doesn't follow from [CRA-1.1] and [CRA-2.1] alone; since, it's not clear what counts as a mental state.

However, to someone who prefers [CRA-1.2] over [CRA-1.1], [CRA-2.2] is gratuitously question-begging in that it simply appends a claim that a state I am in is a brain state. That hasn't been shown. If we strike that offending appositive clause ...

> [CRA-2.2a] The state that plays causal role C = the state I am in when I am performing my causal role.

... we again get a pair of premises that work well together, [CRA-1.2] and [CRA-2.2a]. Again, it seems plausible to theorize that empirical research will eventually show that, when I am performing my causal role, two statements are true: I am in the state of instantiating experiential phenomenon E; and, my brain is in the state of instantiating physical phenomenon P.

However, [CRA-3] doesn't follow from [CRA-1.2] and [CRA-2.2a] alone; since, it's not clear what counts as a mental state. More importantly, it still hasn't been shown that experiential phenomenon E is identical to physical phenomenon P nor that I am identical to my brain; and, both of those must be shown before we can draw an interim conclusion of the form ...

> [IC] The state I am in by virtue of experiencing experiential phenomenon E is identical to the state my brain is in by virtue of exhibiting physical phenomenon P.

Given [IC] and the reasonable assumption that the state I am in by virtue of experiencing experiential phenomenon E is a mental state, [CRA-3] easily follows.

*But how would it be shown that [IC] is true?*

In my view, showing that [IC] is true would require showing that something that exists in an experiencer independent way is identical to something that exists in an experiencer dependent way; and, that doesn't seem possible.

Indeed, as far as I know, no scientists are even working on an attempt to show that something in the brain is nothing other than an appearance to an experiencing I, the subject of the experience. Perhaps this is something experimental philosophers may want to investigate.

If it turns out that no scientists are working on such a project, it calls into question the hope many philosophers apparently have, that scientists will someday show that, for any experiential phenomenon such as the color of an afterimage, there is a physical phenomenon to which it is self-identical. If no scientists are working on such a project, who exactly is expected to fulfill the promises made by promissory materialists?

#### §3.3.3.6.6.1 Closing the Gap by Assuming Causal Closure

I'll start be specifying a couple of examples of causal roles that I claim I play.

Suppose that, while looking at a tomato under optimal viewing conditions, I issue (whether by speaking or writing) a two statement report consisting of

1. $S_1$ - I am now experiencing tomato red.
2. $S_2$ - I take tomato red to be an experiential phenomenon rather than a phenomenal property.

Let us define causal role $C_1$ as the role of commenting on my experience; specifically, the causal role of issuing $S_1$. How much of this role is played by *tomato red*; and, how much is played by me?

Here we have the problem that was mentioned earlier in discussing Papineau's causal argument for materialism. An experiential phenomenon doesn't need a role other than bringing itself to my attention so that I may decide what to do next. Even assuming that such a role is a causal role, the question is whether the outcome, me stating $S_1$, is fully determined by the experiential phenomenon that comes to my attention or whether I have a role in deciding what to report (if anything).

If it is alleged that *tomato red* determines whether I will make a report and what that report will be, I would hasten to reject that position as determinism.

If I have a causal role to play in deciding whether and what to report, I would face the argument from causal closure, ACC. Assuming that the argument does not fall apart upon being stated in the first-person, it seems to go like this:

> [ACC-1] $S_1$, my report concerning my experience, is a physical effect.
>
> [ACC-2] I intentionally initiated the occurrence of $S_1$.
>
> [ACC-3] (therefore) I am the mental cause of $S_1$.
>
> [ACC-4] Any physical$_1$ effect that has a cause has a sufficient, completely physical$_1$ cause.
>
> [ACC-5] (therefore) I am a physical$_1$ cause.
>
> [ACC-6] To be a physical$_1$ cause is to be identical to something physical$_1$ which is a cause.
>
> [ACC-7] (therefore) I am identical to something physical$_1$.

Switching to causal role $C_2$, the issuance of $S_2$, doesn't help. Insofar as that

report is a report, the argument from causal closure applies without alteration.[41]

Problematic for the ACC is that the causal closure principle enshrined in [ACC-4] does not appear to be a finding of physics; and, one may simply reject it on that basis (as I do). In that case, the ACC as I've outlined it above as well as innumerable other arguments for physicalism/materialism may be dismissed as groundless anti-scientific speculation.

Nevertheless, a consideration of the causal role of the experiencing subject concerns the brain/subject relation much more than the brain/experience relation; so, it is an issue to be taken up in §4 after a brief summary of the discourse to this point.

### §3.3.3.7 KA:TNG and Identity Theory, a Summary

In the case of Mary, I assume that tomato red, the particular shade of phenomenal redness that Mary experiences, is an experiential *phenomenon* rather than a property of some kind; and, I show that the choice of linguistic frames of reference has consequences. We have a reasonable expectation that an identity theorist must attempt to show that the experiential phenomenon, *tomato red*, is identical to some physical (presumably neural) phenomenon.

If no such attempt is successful or no such attempt is even made, one may reject claims that defenders of materialism make, such as the claim that Mary only learns an old fact presented in a new way. If tomato red is not identical to NFP-4738 or some other physical$_1$ phenomenon, Mary learns something from becoming acquainted with a phenomenon she had never encountered before.

I formulated a weakened version of the knowledge argument, KA:TNG, which is aimed at theories of physical/phenomenal identity, the only sort of identity theory relevant to Jackson's KA and its variants (including KA:TNG).

Many so-called identity theories either ignore or deny the existence of experiential phenomena by dismissing them as intentional objects, making them eliminative theories rather than theories of physical/phenomenal identity. In considering actual identity theories, I've shown that, when the discussion is cast in terms of phenomena rather than properties, friends of the KA can effectively use the appearance/reality distinction in arguments against physical/phenomenal identity claims.

For that reason, I encourage friends of the KA to shift to a perspective holding that phenomenal redness falls under the concept of phenomenon rather than the concept of property. More generally, I encourage philosophers to insist that first-person phenomenology consists of *phenomena*. To whatever extent a philosophy of consciousness aims to describe and explain first-person phenomenology, it

---

41 Complications would be introduced if I claimed to be the cause of my decision to let experience fall under the concept of phenomenon rather than the concept of property. I reject the allegation that this decision was causally determined by pre-existing neural conditions of which I was unaware. I reject determinism in favor of the assumption that I have the power to choose from among the options open to me at the time of choice; where, *by power to choose*, I mean the power to make choices undetermined by pre-existing conditions of the brain. However, I won't pursue such complications at this time.

must describe and explain first-person *phenomena*.

Along the way, we've found that some identity theories don't demonstrate the identity they claim. Instead, they may describe the circumstances under which philosophers advocating such theories say they will claim that an identity is the best explanation for the evidence. For example, Clark's theory of identity on the basis of parallel phenomenology does not demonstrate an identity. It describes the circumstances under which Clark is willing to make an abductive leap, to infer identity as what Clark considers the best explanation for the parallel phenomenology.

An abductive leap may well be called for if scientists are unable to explain the correlation of experiential and physical phenomena; but, even if an abductive leap is called for, no one is compelled to leap to an inference of identity. One may just as easily leap to an inference of non-identity.

Similarly, some responses to the KA assume an identity without offering an argument for that identity claim. While this defensive strategy may have a rhetorical purpose, defusing an argument against identity theory, it rests on the implicit assumption: if identity theory is true, arguments against identity theory are unsound or invalid. That assumption is certainly true; but, it is no more persuasive than the opposite assumption: if non-identity theory is true, arguments for identity theory are unsound or invalid.

The identity theorist may be aiming for an impasse; but, the non-identity theorist isn't. Non-identity theorist don't need to *assume* non-identity. They have an argument for the *conclusion* of non-identity; namely, that phenomena with different modes of existence can't be identical.

Consequently, the answer to the Churchland/Searle question, [DFQ-1], is that there are two fundamentally distinct kinds of phenomena, physical phenomena and experiential phenomena.

My term for this state of affairs is *phenomenon dualism*.

## *§4 The Brain/Subject Relation*

In considering the brain/subject relation we are faced with choices analogous to those we encountered with respect to the brain/experience relation.

A determined philosopher may conduct a preemptive strike against the possibility of any brain subject relation by denying the existence of the experiencing subject.

If the existence of the experiencing subject is accepted, the next question is whether the brain/subject relation is identity or some non-identity relation. If we admit that the subject of experience exists, but deny that it is identical to something physical$_1$ (the body, the brain, some part of the brain, some part the brain in some state, or whatever); then, depending on your feelings about dualism, we've either achieved or lapsed into dualism.

Naturally, the next question concerns the type of dualism constituted by these affirmations; and, the answer turns on one's position with respect to mental causation. If I, an experiencing subject, claim that I have no causal powers at all, the result is epiphenomenal dualism. If I claim to have causal powers despite not being identical to my brain, I become a property bearer distinct from my brain; and, depending on your feelings about substance dualism, I've either achieved or lapsed into substance dualism. If the experiencing I claims to have causal effects on its brain, we have interactive substance dualism.

In essence, there is a simple argument for interactive substance dualism silently lurking within any discussion of the brain/subject relation. Reduced to its essential claims, the argument for Interactive Substance Dualism, ISD, is:

> [ISD-1] I am, I exist, I occur.
>
> [ISD-2] I am not identical to my brain; to any brain activity with which I am associated; or, to any property of the brain in virtue of which I am, I exist, I occur.
>
> [ISD-3] I have the properties of being able to choose from among the options available to me; and, of being able to initiate action in pursuit of a chosen objective.
>
> [ISD-4] (therefore) I am a property bearer distinct from my brain.
>
> [ISD-5] (therefore) I am a substance distinct from my brain.
>
> [ISD-6] (therefore) interactive substance dualism is true.

Each of [ISD-1] through [ISD-3] is an assumption; and, all are subject to controversy. Together they constitute what I will call *subject causationism*.

The conclusions, [ISD-4] through [ISD-6], follow from prevailing definitions.

By [ISD-2] I am distinct from my brain; and, on the reasonable assumption that having properties makes me a property bearer, by [ISD-3] I am a property bearer; so, I am a property bearer distinct from my brain; hence, [ISD-4].

Given a definition of substance as property bearer, [ISD-5] follows from [ISD-4].

If I'm distinct from my brain and I have no effect on it, I'm an epiphenomenal I; and, the feeling of being able to make choices and initiate activity is an illusion. Consequently, [ISD-3] can't be true unless I'm able to affect my brain as that is my only means of initiating activity; and, that is interactive substance dualism.

In §4.1, I will consider [ISD-1]; the *existence claim*.

In §4.2, §4.3 and §4.4, I will consider [ISD-2]; the *non-identity claim*.

In §4.5, I will consider [ISD-3]; the *claim of causal efficacy*.

## §4.1 Does Experience Presuppose an Experiencing I?

Few philosophers have explicitly denied the existence of an experiencing subject. A few more explicitly affirm the existence of the experiencing subject; but, most philosophers seem to have avoided stating a position on this question (or, so it seems to me).

Searle is well known for holding that subjective states have a first-person mode of existence in that they exist only as experienced by an experiencing I.

> Subjective states have a first-person ontology ("ontology" here means mode of existence) because they exist only when they are experienced by some human or animal agent. They are experienced by some "I" that has the experience, and it is in this sense that they have a first person ontology. (Searle, 2002b, 40-41)

Others philosophers have expressed the same idea. Here's Frege on this topic:

> The field and the frogs in it, the sun which shines on them are there no matter whether I look at them or not, but the sense-impression I have of green exists only because of me, I am its bearer. It seems absurd to us that a pain, a mood, a wish should rove about the world without a bearer, independently. An experience is impossible without an experient. The inner world presupposes the person whose inner world it is. … Things of the outer world are however independent. (Frege, 1956, 299)

And Shoemaker:

> I am of course taking it as an obvious conceptual truth that an experiencing is necessarily an experiencing by a subject of experience, and involves that subject as intimately as branch-bending involves a branch. (Shoemaker, 1996, 10)

The point these philosophers are making – that experiencing presupposes an experiencer – is a key element of any *plausible* theory of the brain/subject relation; but, stating it explicitly suggests that Descartes was right all along about the indubitability of his starting point. If experience presupposes an experiencer; then, as soon as one notices experiencing, one may affirm the existence of the experiencer. Applying this principle in one's own case yields

> **[10] I am experiencing; therefore, I am.**

### *§4.1.1 Am I Mistaken about Being?*

*How could I be mistaken about being when I must be to be mistaken?*

Taking my cue from Hintikka (1965) and Bardon (2005), I realize that certain propositions are verified by their use; and, that among these self-verifying

propositions is “I am something”.

Recall that Descartes depicted the meditating I as reasoning that nothing could make it be the case that I am something when, in fact, I am nothing at all. A malicious demon might fool the meditating I into falsely thinking that I am something specific (a member of the local gentry, say) when, in fact, I am something else (a pauper with delusions of being a member of the local gentry); but, not even a malicious demon of supreme cunning and power could deceive the meditating I into thinking “I am something” when, in fact, I am nothing at all.

Today's meditating I is not beset by malicious demons but must contend with self-eliminating philosophers; otherwise, little has changed since the 17th Century; so, I reply along the lines of Descartes' reply.

When I assert “I am something”, it can not possibly be the case that nothing at all – *nothing* with any mode of existence, *nothing* with any mode of being, *nothing* with any reality of any reality type, *nothing* with any actuality of any kind – *nothing whatsoever* – somehow rose up from the dark abyss of absolute nothingness, ***falsely*** asserted “I am something” and then sank back into the void *without* ***ever*** having been *anything at all*.

It simply makes no sense.

### *§4.1.2 Note on the Usage of Is/Am/Are*

If it can't be the case that I am nothing at all, it must be that case that I am something; and, if I am something rather than nothing at all, I am. It's not possible for “I am something” to be true and, simultaneously, for “I am” to be false.

A moment's reflection also tells me that it's not possible for “I am” to be true and, simultaneously, for “I am something” to be false or for “I am nothing at all” to be true. In view of these conclusions, I will observe the following rule for the use of “am” in first-person statements:

**[Use of Am] I am ↔ I am not nothing at all ↔ I am something.**

And similarly for the second- and third-person cases.

This convention allows us to unify or relate the is of predication and the is of existence. “I am” employs the is of existence; but, “I am not nothing at all” and “I am something” employ the is of predication. Thus, we may conclude that “not nothing at all” or “something” are implied predicates implicit within an assertion of the existential *is*.

### *§4.1.3 Is “I” a Non-Referring Term?*

Elizabeth Anscombe is well-known for holding that “I” is a non-referring term; but, her claim is rather dubious because she limits herself to considering I-statements that may be verified or falsified by external observation; and, she shows exactly how to find the referent of "I" as used in such statements.

Statements like “I am standing” can be verified or falsified from the third-person

perspective.

> There is a real question: with what object is my consciousness of action, posture, and movement, and are my intentions connected in such fashion that that object must be standing up if I have the thought that I am standing up and my thought is true? And there is in answer to that: it is this object here. (Anscombe, 1975, 61)

I concede that the posture of my body determines whether I am telling the truth when I say I am standing. But, how does that show that the "I" in "I am standing" is a non-referring term? Once I have found the object my awareness of action is “connected with” in this manner, have I not found the object to which I am referring when I say "I am standing"?

For example, suppose I say "the Moon is visible tonight" and suppose that I ask myself "What is the object my consciousness is connected with such that that object must be visible if my thought is to be true?". The answer is that it is that object there (pointing to Earth's large natural satellite, Luna, the Moon). By asking and answering my question about the object that is visible, I have found the referent of "the Moon"; and, I see no plausible reason for holding that, in asking and answering her own question about the object that is standing, Anscombe somehow failed to find the referent of "I" as she used it.

In my view, we should concede that "I" is a referring term when used in statement whose truth can be verified by examining the body of the language user. The crucial question is whether "I" remains a referring term when not used in this manner; for example, when it is used to refer to the I as subject rather than as object.

Anscombe concedes that Descartes would have concerned himself with another class of I-statements; for example, "I have a headache", "I am thinking about thinking", "I see a variety of colours" and others. She also acknowledges the difficulty in applying the test of external verification so easily applied to the class of statements she is willing to consider.

> The Cartesianly preferred thoughts all have this same character, of being far removed in their descriptions from the descriptions of the proceedings, etc., of a person in which they might be verified. (Anscombe, 1975, 62)

Nevertheless, Anscombe does not attempt to examine Cartesianly preferred thoughts because they “are not the ones to investigate if one wants to understand 'I' philosophically” (Anscombe, 1975, 63).

One problem with Anscombe's approach is that, to investigate the first-person perspective, we must also consider first-person statements that refer to the experiencing subject – even if only to rule out the value of examining such statements.

Anscombe examines the use of I as subject; but, finds serious difficulties while examining commonly held beliefs about the use of 'I'; for example, the belief that 'I' is guaranteed to refer to its user. The problem as Anscombe sees it is that “if 'I' is a referring expression, then Descartes was right about what the referent was” (Anscombe, 1975, 57). The solution, according to Anscombe, is to declare that "'I' is neither a name nor another kind of expression whose logical role is to

make a reference, at all” (Anscombe, 1975, 59).

*Are we to abandon the inquiry into the use of 'I' as subject because we risk discovering that Descartes was right about the certainty of his starting point?*

I think not. I am reminded of the practice of medieval cartographers of inscribing “Beyond this point there be monsters” at the edge of the known world. Obviously, such warnings did not deter all explorers; and, we are better off for it.

### *§4.1.4 Is there a User of “I”?*

There is a sense of self-contradiction in Anscombe's position. Immediately after stating her solution to the difficulties she points out, she states

> Of course we must accept the rule "If X asserts something with 'I' as subject, his assertion will be true if and only if what he asserts is true of X." (Anscombe, 1975, 59)

In my view, Anscombe is correct about this rule; and, she's correct to suggest that, to apply this rule, we must know whose assertion a given statement is. However, the two examples she gives do not support her claim that 'I' is a non-referring term.

In the case of a translator, we can easily grant that "an interpreter might repeat the 'I' of his principal in his translations" (Anscombe, 1975, 59). But, why would we think that this shows that "I" doesn't refer to anything at all? It seems more plausible to say that we understand that, as stated by the translator, 'I' refers to the person whose words are being translated.

The second example is more difficult. Anscombe imagines someone coming up to her and saying

> Try to believe this: when I say 'I', that does not mean this human being who is making the noise. I am someone else who has borrowed this human being to speak through him. (Anscombe, 1975, 59)

Anscombe points out that such a statement is easily understood despite being taken by the philosopher to be false; but, we can agree with that claim while denying that it shows that 'I' is a non-referring term.

If there actually is a entity in possession of the medium of communication, it's hard to believe that its use of 'I' makes 'I' a non-referring term; although, it seems that this is Anscombe's position.

> If the principle of human rational life in E.A. is a soul (which perhaps can survive E.A., perhaps again animate E.A.) that is not the reference of "I". Nor is it what I am. I am E.A. and shall exist only as long as E.A. exists. (Anscombe, 1975, 62-63)

Let us suppose that it is known for a fact that there are souls; but, that there is a difference of opinion among philosophers and theologians; with some saying that a human being is a human body that has a soul and others saying that a human being is a soul that has a human body. Those holding one perspective might say “I am this living human body” while those hold the other perspective might say “I am the soul that has this human body”.

Clearly, one of those views is false; but, that is beside the point. At issue is

whether any of this shows that 'I' is a non-referring term.

Suppose I say that mathematicians study square circles; and, that you show conclusively that there are no square circles. Clearly, you now have grounds for saying that “square circle” is a non-referring term; but, have you shown that “mathematician” is a non-referring term? No; obviously, you have not shown that there are no mathematicians.

Consequently, we may conclude that reference failure by the predicate term does not entail reference failure by the subject term; and, I will assume that this conclusion applies to first-person self-identifications, statements of the form

**[Identification] I am [a(n)|the|this|that] X**

where the square brackets enclose a pipe (“|”) separated list of optional particularizers and demonstrative terms.

In a statement like “I am an X”, the predicate term performs an ascription (actually a self-ascription) and the subject term performs a reference (actually a self-reference); unless, of course, I is shown to be a non-referring term. So the conclusion I've just reached can be stated this way: A false or faulty self-ascription does not entail a failure of self-referencing.

Now, various groups of philosophers advocate various identifications of what the I is – a body, a brain (or a part thereof), a material mind, an immaterial mind, a soul or a spirit and so on. Each of these claims could be stated as a first-person identification: I am an X.

Suppose that it is eventually discovered for a fact that the I is one of those items and not any of the others. Clearly, one first-person identification will be true and the others false. Equally clearly, one group of philosophers will have been proven right about what the I is. Will that also prove that those philosophers who held mistaken identifications had actually been non-existent all along? No, it will only show that they had been mistaken all along as to what the I is. The mistaken philosophers still have to be to be mistaken.

*What would it take for “I” to be a non-referring term?*

It is not enough to show that “I am an X” is false for some values of X. It would be necessary to show that I am nothing at all – that there is nothing to which any predicate may be attributed. Yet anyone asserting “I am nothing at all” thereby shows why that statement is false: something made the assertion. We have no reason to think that no one may refer to that something; or, that no may self-identify as that something.

The “I” in “I am something rather than nothing at all” refers to that which asserts “I am something rather than nothing at all”.

Alternately, I am the user of “I” whenever I say “I am the user of 'I'”.

Something is using the term “I” for the purpose of referring to the user of “I”. There has to be such a user; such assertions to not bubble up from the void of nothingness the way virtual particles emerge from the quantum value to exist for too brief a time to be detected.

We may disagree as to what that I is; but, unless you can defend your assertion of “I am nothing at all”, you have no basis for concluding that “I” is a non-referring term.

### *§4.1.5 Who is the User of “I”?*

Having concluded that there is a user of “I”, it seems very natural to wonder what the user of “I” may say about itself. However, before moving on to ask what the I is, I want to amplify on a point previously made – that Anscombe was correct to point out that, as spoken by a translator, “I” may refer to the person whose words are being translated rather than to the person doing the translating.

In view of this, we need to recognize exceptions to the rule that “I” refers to its user; or, we need to be more flexible as to who is the user.

Particularly when speaking about assertions, it is natural to assume that only a speaker or a writer can be the user of “I”. For example, when I wrote “I want to amplify on a point previously made” I intended to refer to me, the author of the document you are reading.

However, I've written other statements into this document intending the term “I” to be used by you, the reader, for the purpose of self-referencing. Consider the following first-person statement.

**[FPS-1] I am reading this sentence.**

When you are evaluating the truth of [FPS-1], you should 'use' this sentence by letting its token of “I” refer to you while you go about your task of discovering whether [FPS-1] is true of *you* while you are reading it, *not* whether it is true of me while you are reading it.

I call this the *reader referencing* use of “I”. One might also call this the fellow traveler or co-meditative use of “I” out of respect to Descartes who expected his readers to meditate along with him when they read the first-person prose of his Meditations. I do not hold his belief that everyone who thinks along with me in the first person will reach the conclusions that I reach; but, I am making a first person argument in which some of the propositions are stated using the reader referencing use of “I”.

In particular, [10] is so stated.

*Why use reader referencing first-person propositions?*

As you evaluate the truth of “I am experiencing; therefore, I am”, the question at issue is whether that statement is true of *you*. For all you know, I could be a terminator robot sent back in time by dualists of the future for the purpose of attacking materialism for its blind faith in the causal closure principle.

Less flamboyantly, it is essential to read [10] as using a reader referencing “I” because, from the from the third person perspective, I and my zombie twin are indistinguishable, as was pointed out by Chalmers (1996, 198). From the first-person perspective, experience is self-evidencing. I am directly acquainted with the experiential phenomena that constitutes *my* experience; consequently, I may

conclude that I am not a zombie. But it is only from the first-person perspective that I know that I am not a zombie. From the third-person point of view, all we know is that we both have brain activity.

## §4.2 What Am I?

Having concluded *that* I am, I naturally move on to the next question: *what am I?*

Curiously, as I ponder this question, I gradually realize that just knowing *that* I am does not tell me *what* I am; at least, not immediately. I will have to work for whatever conclusion I reach.

Some obvious efforts are unhelpful. From the fact that I am plus the convention that whatever is exists, it follows that I am an existent of some sort; but, this conclusion is not very informative.

Biographical information is also unhelpful. It is a fact about me that I am an American; so, the statement “I am an American” is an informative, first-person, self-identification; but, it is irrelevant to the task at hand. I want to know *what* I am, not *who* I am.

After pondering my predicament, I realize that it defines this stage of the epistemological journey.

**[11] I know *that* I am; but, not *what* I am.**

I might advance a provisional answer, an *initial* identification of what I am based on what I know so far: that I am an I (whatever that may turn out to be); and, that I experience. From these two statements it seems reasonable to conclude

**[12] I am an experiencing I.**

Such an identification[42] does not fully answer the “*what am I?*” question; but, at the beginning of the epistemological journey, I don't expect to have a final, fully satisfying answer.

At this point in his epistemological journey toward greater understanding, Descartes adopted a similar identification, “I am a thinking thing”. However, that wasn't the end of the matter for Descartes. He wanted to know whether the thinking thing is a body or a soul. He worked long and hard for his answer; but, his work was flawed.[43]

Suppose that, upon reflection, I decide to extend [12] by individuating the

---

42 A number of alternate identifications would be possible at this point, including “I am an experiencer”, the demonstrative identifications, “I am this experiencing I” and “I am this experiencer”, as well as versions with “the” in place of “this”.

43 Even among those, such as myself, who say that Descartes found an indubitably true starting point, opinions may vary as to where Descartes went wrong. In my view, the great Cartesian Circularity begins early in the Third Meditation when he invokes the natural light (a term which seems to refer to the capacity for intuition construed as a divine gift) in the argument for the existence of God who is invoked to guarantee the accuracy of the natural light. Further, one may recognize that humans have an intuitive ability to recognize truth apart from the logic by which it is made provable; but, one may nevertheless deny Descartes' expectation that everyone who meditates along with him will intuit the same truth.

experiencing I according to its stream of experiences, yielding

**[13] I am the experiencing I of this stream of experiences.**

While I accept that you, other humans and at least some animals also have an associated stream of experiences, I only have access to *this* stream. Experience has a sense of mineness to it; so, [13] could be restated as

**[13.1] I am the experiencing I of my stream of experiences.**

To link my terminology with a more traditional third-person description of the first-person perspective, I will identify the experiencing I and the experiencing subject of experience.

**[14] I am the experiencing subject of my stream of experiences**

or, more abstractly:

**[14.1] An experiencing I is the experiencing subject of its stream of experiences.**

Even given the self-identification as the subject of experience, I don't know whether the subject of experience I am is identical to the brain I seem to have or to an immaterial entity (a mind/soul of the Cartesian sort) or neither. I might arise from or be generated by the brain or by an immaterial entity without being identical to it. It might require both the brain and an immaterial mind to generate an experiencing I. Those are the possibilities that I can envision; but, the true state of affairs might be stranger than I can currently imagine.

I might not *ever* be able to make a positive identification of an informative sort; and, I take Kant, the First Mysterian, to have made a strong argument in his discussion of the paralogisms in the first Critique that one can not decide this point by rational means alone. What we might be able to conclude by rational means together with the findings of empirical research is an open question; but, that's not my concern at the moment.

At the moment, my claim is only that we have sufficient grounds for rejecting brain/subject identity.

## §4.3 What Am I *Not*?

I notice that I am conscious. After careful consideration, I conclude that this is a new self-evidencing fact. To paraphrase Searle, if it seems to me that I am conscious, I am conscious. (Searle, 1997, 213)

Asserting “I am conscious” prompts a further question, “A conscious what?”. From the fact that I am and the assumption that whatever is is a being of some sort, I may conclude that I am a being of some sort. Combining this conclusion with the fact that I am conscious yields the conclusion that I am a conscious being.

It may be objected that, by combining an obvious fact (that I am conscious) with an uninformative fact (I am a being of some sort if and only if I am not nothing at all), I have accomplished little. I reply that I've laid the foundation for engaging the work of philosophers who wrote about conscious beings; because, I now

know that I am an instance of the class of entities about which they wrote.

David Barnett, for example, has this to say about conscious beings:

> I argue that, unlike your brain, you are not composed of other things: you are *simple*. My argument centers on what I take to be an uncontroversial datum: for any pair of conscious beings, it is impossible for the pair *itself* to be conscious. Consider, for instance, the pair comprising you and me. You might pinch your arm and feel a pain. I might simultaneously pinch my arm and feel a qualitatively identical pain. But the pair we form would not feel a thing. Pairs of people *themselves* are incapable of experience. Call this *The Datum*. (Barnett, 210, 161)

The Datum seems to be intuitively obvious. If I take two conscious beings chosen at random and arbitrarily declare them to be a pair, I do not thereby bring into being a third experiencing subject, the pair itself.

Barnett then asks *why* the pair would not itself be a conscious being. He rejects the possibility that the pair does not itself experience because it has an insufficient number of parts. It wouldn't help if we started composing triplets or even larger collections of conscious beings. The collection will not itself be conscious.[44]

Barnett then goes on to reject other possible explanations as equally insufficient: that the pair is not a conscious being because it has an insufficient number of parts, or parts of the right kind or parts arranged in the right structure or standing in the right relations to each other. That leaves him with only one plausible alternative: that a pair of conscious beings is not itself a conscious being because it is not a simple thing, a thing without parts.

Barnett's reasoning appears to be sound; so, I conclude that I am a simple thing. Now, it goes without saying that the brain is a complex entity having billions of neurons; hence, it can not be a simple thing.

It follows that I am not identical to my brain.

> **[15] I am not identical to any brain activity with which I am associated; nor, to any properties of the brain in virtue of which I am, I exist, I occur.**

## §4.4 Where Am I?

In the movie, *Fantastic Voyage*, a great scientist was gravely injured in an assassination attempt. The damage to his brain was inoperable … from the outside, anyway. So, with impeccable movie logic, the response team shrank a small submarine and its crew of 5 and injected it into the bloodstream of the patient. After a perilous journey, the sub made its way to the patient's brain and the intrepid neuronauts repaired the damage from the inside.

---

44 Barnett points out that a stipulation to this effect was written into the foundations of functionalism by Putnam. “No organism capable of feeling pain possesses a decomposition into parts which separately [are capable of feeling pain]” (Putnam, 1967, 227). Putnam's purpose was to rule out the possibility that a swarms of bees could be considered a distinct experiencer of pain no matter how functionally organized; but, he did not draw the conclusion that Barnett will draw.

Suppose we dispense with the silliness of shrinking people and imagine the possibility of building a microscopic drone submarine equipped with various sensors and the means for two way communication with a controller, who could easily be the subject of the experiment, the philosopher into whose body the submarine is injected.

I assume command of the Nanobotic Neuro-Submarine, NNS *Enterprise*!

*What do I do?*

For a shakedown cruise, I might have the sub injected into my bloodstream and navigate to the visual cortex in search of afterimages. When the sub is in position, I stare at a tomato as if I had just been released from monochromatic confinement. I look away and see an otamot green afterimage.

While enjoying the afterimage, I order the Enterprise to conduct a sensor sweep of its surroundings. I can detect many neurons (which are much larger than the NNS *Enterprise*); and, I can detect the chemicals by which they signal each other. With other instruments, I can detect electrical activity. But, I do not find a single instance of *otamot green* in, on or hiding behind some neuron.

The Enterprise relays the raw data to my desktop supercomputer which pieces together a topological map of neural firing patterns, some involving millions of neurons; but, I do not find an instance of *otamot green* within this electrical activity.

I order the Enterprise to probe the cytoskeleton of a nearby neuron; and, it deploys a probe at the end of a carbon nanotubular cable, snakes it through a convenient cellular receptor site and attaches it to a suitable microtubule. Using weak measurement techniques, the probe gently queries quantum phenomena traveling along the microtubule without provoking any collapse events.

Still, I do not find an instance of *otamot green*. I find no afterimages, no experiential phenomena and no qualia of any kind lurking withing the microtubules of my brain.

*Would it have helped to have the assistance of other philosophers?*

Perhaps I could inject myself with several NNS and invite other philosophers to control all but the Enterprise. We could travel around as a neuro-naval task force. We could go to the visual cortex and stimulate a neuron. Each sub commander would be able to detect the resulting electro-chemical activity because it is an objective, physical$_1$ phenomenon. I alone would experience a color quale. The experiential phenomenon I experience but the operators of the other NNS sailing around my head do not experience only exists subjectively (in my experience) rather than objectively (in my brain).

Following Dennett, I conclude that nothing in my brain is an instance of otamot green.

*Now what?*

I remind myself that the sub and its equipment appear to have functioned within expected parameters. So, despite the negative results of my search for afterimages, the purpose of the shakedown cruise has been accomplished. The

NNS *Enterprise* has been proven seaworthy.

I realize that, if I am unable to find an afterimage, I'm not likely to find the experiential phenomena associated with my other senses. But, it occurs to me that there may be something else in the brain of interest to philosophers of consciousness: the experiencing I – *this* that I am. So, I decide to take the Enterprise on a voyage of self-discovery.

Where do I go?

I suppose that I could go to the pineal gland; but, I'm not looking for animal spirits. I'm looking for this experiencing I – that which is looking for itself. Although I have concluded on rational grounds that I am not identical to any brain activity, in the spirit of scientific research, I should test the null hypothesis: that I am identical to a group of neurons. I decide to look for a group of neurons that denies being a group of neurons.

Recent scientific research suggests that heightened activity in the Precuneus, a region within the parietal cortex, is associated with “self-processing operations, namely first-person perspective taking and an experience of agency” (Cavanna and Trimble, 2006, 564); so, I decide to go there.

As I get underway, I turn on the sub's virtual reality mode. telemetry from the sub is displayed via the VR headset I am wearing. The motion of my hands as I push buttons in the VR display is intepreted by the gloves that I am wearing and relayed to the Enterprise which responds accordingly. The experience seems like being there on the bridge of the NNS *Enterprise*.

I arrive in the Precuneus. Once again, I find electrochemical activity; but, despite making detailed measurements of various neural firing patterns, I find no evidence that these neurons are me denying that I am them.[45] Indeed, I see no evidence that the neurons involved in this activity either affirm or deny being the neurons involved in this activity.

Perhaps the problem is that our technology just isn't advanced enough to know what a group of neurons affirms or denies. Perhaps, we don't know the code by which neurons process the information associated with such claims.

If so, I may have to defer my attempt to discover what I am by discovering where I am. I decide to aim for a lesser target. Perhaps, given recent scientific discoveries, I might find myself in the act of taking action.

Leon Gmeindl and his team note that

> Functional magnetic resonance imaging (fMRI) studies have indicated that the medial superior parietal lobule (mSPL), a component of the dorsal attention network, is activated during shifts of attention between spatial locations, features, objects, sensory modalities, task sets, and working memory representations. (Gmeindl et al., 2016, 2176)

Prior studies involved cued shifts of attention; and, the team wanted in

---

45 As a control for experimenter bias, I encourage identity theorist to install their own nanobotic neuro-submarines and search their brains for evidence that some group of neurons is me affirming that I am them.

investigate spontaneous shifts of attention to avoid complications from the effects of cuing. They also dispensed with the need for subject to indicate behaviorally (e.g. by pushing a button) when a shift took place.

Gmeindl and team found that "During these uncued, self-generated shifts of attention, mSPL was transiently active, consistent with the hypothesis that it plays a role in reconfiguring attention. (p. 2183)

The researchers also found that activity in the right middle frontal gyrus (rMFG) increased

> … prior to self-generated attention shifts, and it increased earlier than the activity associated with cue-driven attention shifts. This finding suggests that rMFG participates in the preparation to reorient attention ... We also observed an early increase in dACC [dorsal anterior cingulate cortex] activity that was associated with self-generated, but not cue-driven, attention shifts. (p.2183)

The researchers conclude

> Together, these findings suggest that medial frontal cortex activity reflects a common locus of early-stage processing that gives rise to self-generated behavior, including both overt action and covert orienting of attention. We suggest that the dACC and rMFG are core components of the brain network underlying the will to act. (p.2183)

Now, I have a plan. I will go to each of these regions in turn and investigate. I will park the Enterprise and program it to take readings of the target neural structures. My plan is then to make a decision and see if there is any activity that correlates with the sense of agency.

To duplicate the condition that Gmeindl and his team investigated. I decide that I will move my eyes from left to right and back at self-chosen intervals. I don't want what I see to affect the outcome; so, I set up duplicate monitors on which to display the telemetry from the Enterprise, one more to my right and the other more to my left. I may shift my eyes as I wish; but, each monitor will display the same information as the other monitor.

Let us suppose that the results are what we might predict on the basis of Gmeindel et al. There is activity in the rMFG and the dACC prior to activity in the mSPL.

*What might I conclude from this fact, (assuming it turns out to be a fact)?*

Should I conclude that I am identical to one of these regions in my brain or to activity in one of these regions? Since I will certainly continue to deny being them, would they not have to deny being self-identical (in addition to performing their other functions in bringing about a willed event)?

Indeed, any group of neurons considered a candidate for being me would have to have this feature: they must somehow deny being themselves. Perhaps there is a special group of neurons that have only the function of affirming or denying self-identity. Or, perhaps there is a special group of neurons that have this function in addition to whatever else they do.

In any case, let us assume that the difference between the identity theorist and the non-identity theorist is that the identity theorist affirms (for some group of

neurons) "I am that group of neurons" whereas the non-identity theorist says "I am not that group of neurons" (for any group of neurons). Let us assume further that this difference is a mental difference. If there is no associated physical difference, there is a mental difference without a physical difference; and, supervenience physicalism fails.

Setting aside the dispute between identity theorists and non-identity theorist, how might we build on research of this sort?

It would seem that further research is needed to connect activity in any of these regions of interest with the eye muscles that actually move my eyes. From the first person perspective, it seems to me that, during an eye moving event that seems voluntary to me, a shift in my attention is followed by an eye movement corresponding to my shifted attention.

The eye muscles are a long way from the mSPL/rMFG/dACC complex. Is there an output signal from this complex to the eye muscles; or, are their actions mysteriously coordinated through the miracle of gamma synchrony or by some other means? No one knows.

Since I claim the ability to move my eyes, I would expect some signal to originate with me, whatever it is that I am. I might be or I might act upon one of the regions in which preparatory activity has been found. So, more research is also needed concerning the input signal.

While we await further bulletins from the frontiers of science, there is a philosophical issue that must be dealt with. It seems that many philosophers simply deny having the ability to move their eyes voluntarily.

I am writing this document with my editor in split screen mode. The passage I'm writing is in the right side pane. My notes and source material are displayed in the left side pane. As I type, I frequency feel the urge to refer to my notes ... and I find myself looking at the left pane.

I don't know which eye muscles moved or how they moved to bring about that result. I don't know which nerves stimulated which eye muscles. From this first person perspective, it seems as if intending to look at my notes is followed by an appropriate eye movement – appropriate because it implemented my intent.

However, many philosophers would dispute this interpretation. In their view, there are no voluntary actions. All actions are determined by brain activity which also generates an illusion of agency for the experiencing subject. So, if neuroscientists ever want to help resolve the philosophical dispute over the reality of voluntary action, it would help to repeat some fMRI studies of voluntary actions while look for evidence of activity in areas of the brain associated with generating illusions, delusions and hallucinations.

Until such research is conducted, we must continue to debate the issue of causal efficacy on the basis of rational arguments. And, to that topic I now turn.

## §4.5 Am I Causally Effective?

Having affirmed [ISD-1] and [ISD-2], I have arrived at the banks of the Rubicon.

Affirming [ISD-3] seems to make me a property bearer; hence, a substance. Since I have already denied being identical to my brain, I would be a substance distinct from my brain; hence, substance dualism would be true.

Denying [ISD-3] seems to make me an epiphenomenon.

Being forced to choose between such controversial alternates poses quite a dilemma for any philosopher; but, refusing to affirm or deny [ISD-3] makes one's philosophy of consciousness incomplete.

It is not even clear that the question of whether [ISD-3] is true or false can be decided by rational argument. It may be that the philosophers are condemned to assume their way across the Rubicon.

### *§4.5.1 The Liquidity Predicament*

To better understand the situation of an experiencing I who denies being identical to its brain or to some pattern of brain activity, let us consider Searle's attempt to describe the relation between consciousness and its neuronal base.

> Consciousness is caused by lower-level neuronal processes in the brain and is itself a feature of the brain. Because it is a feature that emerges from certain neuronal activities, we can think of it as an 'emergent property' of the brain. An emergent property of a system is one that is causally explained by the behavior of the elements of the system; but, it is not a property of any individual elements and it cannot be explained simply as a summation of the properties of those elements. The liquidity of water is a good example: the behavior of the $H_20$ molecules explains liquidity but the individual molecules are not liquid. (Searle, 1997, p. 17-18)

As a phenomenon dualist concerned with distinguishing my position from property dualism, I deny being a *property*, emergent or otherwise. Instead, I claim to be an emergent *phenomenon.* But, nothing in the use I make of the liquidity metaphor turns on the distinction between emergent properties and emergent phenomena.

For present purposes, I'll assume that the liquidity metaphor adequately models the concept of emergence; meaning, at minimum, that an emergent phenomenon is not identical to that from which it emerged. We would not normally say that something emerges from itself; so, intuitively, this makes perfect sense; and, the model is consistent with the conclusion of non-identity reached previously.

Suppose that I am sitting here at my keyboard typing away when I suddenly realize that I (whatever I may turn out to be) have somehow emerged from my neurobiological correlates (whatever they may turn out to be). After taking stock of my situation, I assert “I am conscious!” with an appropriate sense of wonder.

Now, because of the assumption of non-identity built into the concept of emergence, I already know that I am not identical to the neural correlates from which I have emerged. But, the liquidity metaphor is suggesting a further conclusion; namely, that I am conscious *whereas none of my neural correlates are themselves conscious*.

My situation has become a predicament.

In asserting that I am conscious whereas my neurobiological correlates are not conscious, I ascribe the predicate “conscious” to *this* (whatever this is that I am) but not to my neurobiological correlates to which I am not identical.[46]

In ascribing this predicate, have I detected a corresponding property that I have? I would say “No”.

In my view, it just happens to be the case (in English, anyway) that the name for a one-argument predicate is “property”. If I self-ascribe a one-argument predicate, I have self-ascribed a property in that sense of the term. But, I deny having thereby proven that I am a property bearer (where “property bearer” is the meaning of “substance”) in the philosophical sense.

It is merely a quirky, contingent fact of the English language that the name of a one-argument predicate is “property”. We can easily imagine a language in which “substance” means “property bearer” just as it does in English but in which the name of a one-argument predicate is “zargon” while the usage of “property” is otherwise unchanged. In such a language, even after I self-ascribe a predicate, the question of whether I have a corresponding property remains open.

For these reasons, I conclude that predicate ascription does not entail property detection.[47] This conclusion seems intuitively correct. It should take more than admitting to being conscious to show that substance dualism is true.

However, that leaves me with the problem of determining whether statements I take to be true of the experiencing I – that I experience making choices and initiating actions – provide evidence that I have a property that *would* make me a property bearer and, hence, a substance.

*What more than predicate ascription is required to make me a property bearer?*

In my view, the circumstances of predicate ascription would have to be such as to support the conclusion that I have a property corresponding to the predicate ascribed. Fortunately, there is at least one situation which satisfies this condition, claims of agency.

If my experience of agency is veridical at least some of the time, I, this consciousness, have an impact on the ($physical_A$) world; and, it would follow that I have the property of being able to cause a certain range of ($physical_1$) effects.

## *§4.5.2 What Would Make Agency Possible?*

Searle (2001; 2006) investigated the conditions that would make freedom of will

---

46 To be a bit more precise than I think is required at the moment, I should say that my neurobiological correlates are not conscious in the same sense in which I am conscious. I don't object to saying that my brain is conscious in the same sense in which I say that my brain undergoes a reaction to incoming stimuli. All of that can happen without any consciousness on my part.

47 This principle should be uncontroversial. It is quite possible to admit that a statement like “that tomato is red” is true (in a manner of speaking) while also denying color realism. The predicate is ascribed; but, the tomato lacks a corresponding property.

possible. The conclusion of this transcendental argument is that “the condition of possibility of the adequacy of rational explanations is the existence of an irreducible self, a rational agent, capable of acting on reasons”. (Searle, 2006, p. 57)

Ah! Now *that* would be me. I am one of *those!*

A bit earlier in the same work, Searle made it clear that I couldn't be just a bundle of perceptions, a mere Humean I or self.

> We experience ourselves acting as rational agents, and our linguistic practice of giving explanations reflects the gap (because the explanations do not cite causally sufficient conditions). And, for their intelligibility, these explanations require that we recognize that there must be an entity -- a rational agent, a self or an ego -- that acts in the gap (because a Humean bundle of perceptions would not be enough to account for the adequacy of the explanations). The necessity of assuming the operation of an irreducible, non-Humean self is a feature both of our actual experience of voluntary action and the practice that we have of explaining our voluntary actions by giving reasons. (Searle, 2006, p. 56)

However, at this point in his argument, Searle is only trying to uncover the conditions for the possibility of experiencing a sense of agency – he hasn't concluded that those experiences are veridical. So, the interim conclusion of the argument to this point is very modest – that being a causally effective agent presupposes being the non-Humean self just described.

Searle now goes on to state the problem:

> The problem of free will is whether the conscious thought processes in the brain, the processes that constitute the *experiences* of free will, are realized in a neurobiological system that is totally deterministic. (Searle, 2006, 61)

Searle considers two hypotheses concerning the relation between the neurobiological conditions at the time we experience making a choice and the outcome, the action taken.

> Hypothesis 1: The neurobiological conditions are causally sufficient to determine the outcome.
>
> Hypothesis 2: The neurobiological conditions are not causally sufficient to determine the outcome.

On Hypothesis 1, our feelings of deliberating while apparently making what appears to us a choice from among the options available to us, our feelings of having made a choice, our feelings of initiating action in pursuit of our chosen objective – they're all epiphenomenal illusions.

> [The apparent effort of] making up our minds is simply not a causally relevant aspect in determining what actually happens. Our decision was already fixed by the state of our neurons even though we thought we were going through a conscious process of making up our minds among genuine alternatives, alternatives that were genuinely open to us, even given all of the causes. (Searle, 2006, 68)

Rejecting Hypothesis 1 avoids epiphenomenalism; but, accepting Hypothesis 2 raises the question of how to explain how Hypothesis 2 is possible. Searle's analysis uncovered three conditions necessary for the truth of Hypothesis 2.

Those conditions are (Searle, 2006, 71-73):

1. Consciousness, as caused by neuronal processes and realized in neuronal systems, functions causally in moving the body.
2. The brain causes and sustains the existence of a conscious self that is able to make rational decisions and carry them out in actions.
3. The brain is such that the conscious self is able to make and carry out decisions in the gap, where neither decision nor action is determined in advance, by causally sufficient conditions, yet both are rationally explained by the reasons the agent is acting on.

In condition 1, the appositive clause “as caused by neuronal processes and realized in neuronal systems” seems unnecessary and question begging; and, I would strike it entirely. There is no way a philosopher cogitating in his armchair can rule out the possibility that consciousness emerges from a sub-neuronal level, as is advocated by the Penrose/Hameroff theory that consciousness is generated by quantum computations at the sub-neuronal level, in the microtubules of the brain.

In any case, Searle admits that we don't know how the brain could satisfy these conditions; and, argues that a theory explaining how the brain does it would have to incorporate quantum indeterminism (but not quantum randomness).

> The indeterminacy at the micro level may (if Hypothesis 2 is true) explain the indeterminacy of the system, but the randomness of the micro level does not by itself imply randomness at the system level. (Searle, 2006, 76)

Searle closes almost apologetically.

> Hypothesis 2 is a mess, because it gives us three mysteries for one. We thought free will was a mystery, but consciousness and quantum mechanics were two separate and distinct mysteries. Now we have the result that in order to solve the first we have to solve the second and invoke one of the most mysterious aspects of the third to solve the first two. (Searle, 2006, 77-78)

### *§4.5.3 The Searlean Dilemma*

If I am a non-Humean self such as that which Searle believes is necessary to make our experience of agency veridical, a rational agent distinct from my neural correlates but somehow able to initiate intentional actions anyway, it seems to follow that I have the properties of being able to choose from among the options available to me and of being able to initiate actions to achieve my chosen objective. If I have these properties I can hardly deny being a property bearer. Since I've already claimed to be distinct from the brain with which I am associated, I would be a property bearer distinct from my brain which, of course, has properties of its own. Given that one meaning of “substance” in the philosophical lexicon is “property bearer”, I would be a substance distinct from my brain.

Such a state of affairs would constitute substance dualism. Indeed, since we arrived at this point by attempting to defend free will and the possibility of volition (intentional actions, mental causation), this state of affairs would

constitute interactive substance dualism; although, not necessarily of the Cartesian variety.[48]

Of course, Searle does not agree that his views constitute interactive substance dualism whether of the Cartesian variety or some other. He denies being a dualist of any kind; and, as we saw when reviewing his defenses, he tries very hard to avoid being seen as such.

In the section on the brain/experience relation, I considered several of Searle's defenses that seemed appropriate to consider in the context of an inquiry into the nature of that relation. Searle also denies that his position on free will makes him a dualist. Searle asserts that, despite being *ontologically* irreducible to brain activity, consciousness (the experiencing I) is *causally* reduced; meaning, that "consciousness has no causal powers beyond the powers of the neuronal (and other neurobiological) structures." (Searle, 2006, 50)

This position is highly problematic. First, it is hard to distinguish being causally reduced from being an epiphenomenon. I will examine this question in §4.5.3.1.

Next, in §4.5.3.2, I will examine what seems to be a conflict between the thesis of causal reducibility and the claim of rational agency.

If the causal reduction claim is untenable because of its epiphenomenalism, Searle needs a better defense to the charge that his position on free will constitutes interactive substance dualism. In §4.5.3.4, I will consider the causal closure principle and show that it is not a viable principle, whether deployed as a defense to interactive substance dualism or for some other reason.

Finally, beginning in §4.5.3.5, I will show that, absent the causal closure principle and the thesis of causal reducibility, Searle's analysis of free will has some support among physicists.

### §4.5.3.1 None Dare Call it Epiphenomenalism

Searle considers consciousness to be a high level or emergent property whose causal powers are fully accounted for by the causal powers of its neuronal base. He then argues that we're not required to consider this state of affairs to be epiphenomenalism.

> Nobody thinks that we are forced to postulate that solidity is epiphenomenal on the grounds that it has no causal powers in addition to the causal powers of the molecular structures, nor do they think that if we recognize the causal powers of solidity we are forced to postulate causal overdetermination, because now the same effect can be explained either in terms of the behavior of the molecules or the solidity of the whole structure. (Searle, 2002a, 62)

But, surely, this is empirically false; for example, William Seager thinks what Searle claims nobody thinks.

---

48 See Nida-Rümelin (2006) or Lowe (2006) for examples of non-Cartesian substance dualism consistent with Searle's position minus the causal closure principle and the causal reducibility thesis.

> ... high level features exactly fulfill the traditional conception of epiphenomena: they are features dependent on the underlying causal structure which add nothing to the causal powers of the world. (Seager, 2006, 31)

Thus, Seager accepts Searle's claim that consciousness is a high level, emergent feature of the world; but, he does not exempt it from being an epiphenomenon.

> Within the [Scientific Picture of the World] SPW, conscious awareness does not seem to be a basic feature of the world but rather another high level feature, one that involves the mass action of at least millions, if not hundreds of millions, of neurons. But if we accept that consciousness is another high level feature, we must conclude that it, like all other high level features, is causally impotent. Consciousness does not participate in the 'go' of the world; it does not add any constraints upon state evolution that are not already present because of the fundamental physical features. That is, we must conclude that consciousness is epiphenomenal. (Seager, 2006, 35)

In my view, Searle needs a better reason for claiming that his view does not constitute epiphenomenalism, which is suspect in the view of many philosophers, including Searle who calls it an “absurd consequence” of dualism (Searle, 2007a, p. 175). Furthermore, avoiding epiphenomenalism is a key component of Searle's attempt to show that his view on free of will does not constitute dualism.

> On standard versions of dualism it is hard, if not impossible, to see how consciousness could have any causal impact in the world, yet we know that it does have a causal impact: I decide to raise my arm, I form a conscious intention-in-action to raise my arm, and then the arm goes up. There isn’t any doubt that my conscious intention causes the arm to go up. (Searle, 2007a, p. 175)

Searle's argument seems to be that dualism implies epiphenomenalism; that his position does not imply epiphenomenalism; hence, his position is not dualism. Consequently, if Searle fails to distinguish the consequences of his position from those he attributes to dualism, his argument fails.[49]

*Does Searle's position constitute epiphenomenalism?*

Consider a slight modification of Searle's own example. I want to raise *an* arm. I deliberate as to which arm to raise. I decide to raise my right arm; and, my right arm goes up. I have no doubt that my conscious intention caused my right arm to go up. Now, it *seems* to me that I have causal powers, the power to choose which arm to move; and, the power to move the arm I chose to move. But, at this point, I don't yet know whether what seems to be the case actually is the case.

At this point, I would certainly choose Hypothesis 2 over Hypothesis 1 because it is consistent with claims of subject causation: that I may initiate action when the causal conditions in my brain are insufficient to determine the action to be taken. However, if I now introduce the thesis of causal reducibility, it nullifies the benefit of Hypothesis 2. If I lack any causal powers beyond those of my brain, it becomes impossible for me to initiate any action when the causal powers of my brain are insufficient to determine the action to be taken.

---

49 There are other possible objections. As far as I can tell, most dualists would agree with Searle's claim that dualism makes explaining interaction difficult; but, that merely acknowledges the profound mystery we are all investigating. Few actually embrace epiphenomenalism. Is there a non question begging way to show that the first premise is true? If not, the first premise would be quite dubious.

Clearly, the thesis of causal reducibility substitutes an epiphenomenal I for the causally effective subject; and, should be rejected on that basis.

While we don't know whether Hypothesis 1 or Hypothesis 2 is true, it is clear that Hypothesis 1 results in epiphenomenalism and that Hypothesis 2 plus the thesis of causal reducibility also leads to epiphenomenalism. So, Searle has failed to distinguish the consequences of his views from the consequences of dualism; so his argument for not being considered a dualist also fails.

### §4.5.3.2 The Conflict with Rational Agency

Suppose I accidentally touch a hot surface. The signal from nerve endings in my hand triggers a reflex action before it even reaches my brain. In such cases, I, consciousness, don't even try to take credit for moving my hand. I'm content to say that my brain moved my hand *all by itself*.

The mystery of intentional action arises from the intuition that not all bodily movements are reflex actions. Some of my actions appear to be intentional; meaning, that they appear to be initiated by my intent to achieve an effect.

Someone who tries to avoid interactive substance dualism by denying that the experiencing I has causal powers may also try to avoid epiphenomenalism by claiming that the experiencing I may take credit for the actions of its brain.

*Under what circumstances may the experiencing I take credit for the actions of its brain?*

There are circumstances in which some sort of dual control is possible. For example, it is known that the eyes are constantly moved by the nervous system; and, one may reasonably infer that the brain uses this information along with information passed on by the retina to generate a stable visual field.

Now, it seems to me that sometimes my eyes move because I've intended to look at something I wasn't previously looking at. Suppose that while typing this sentence I decide to look at the print of a landscape scene hanging on the wall to my left. My head rises slightly and my eyes move. The print is further from me than the screen at which I had been looking; so, I assume that my eyes were moved so as to focus accordingly.

*Did I move my eyes?*

In an overly strict sense, I could say that I did not move my eyes. I could say that I intended to look at the landscape scene and my brain moved my eyes as necessary to implement my intention. I can rely on my brain to do things like that because it is intent-responsive; at least, to some extent. However, in a looser but more appropriate sense, I can say that I moved my eyes because I'm willing to take credit for the actions of my brain when I initiate those actions.

*Did I initiate the action?*

I have before me a book containing autostereograms, flat two dimensional images into which the artist has embedded depth information. If the viewer focuses on a point farther away from the eye than the page on which the image is printed (wall-eyed vision), a 3-D image will emerge rising up from the page.

Usually, the viewer can see a second 3-D image by focusing the eyes on a point closer to the eye that the page (cross-eyed vision); but, this image will be a reversal of the wall-eyed 3-D image. Height will become depth.

I opened the book at random and my eyes auto-focused on the 2-D image printed on the page. I can continue to look at the 2-D image for any length of time as I ponder the question of whether I will attempt to see the wall-eyed 3-D image or the cross-eyed 3-D image. It seems obvious to me that I'm in a situation in which the causal conditions in the brain are not sufficient to determine the action to be taken. It seems to me that I have to make a choice.

Suppose that I form the intent to see the wall-eyed 3-D image. I know from past experience that I have to relax my eyes. I form the intention to do so and my eyes move accordingly. I presume that my brain is coordinating the movements of the various eye muscles involved; I'm certainly not aware of doing so. As my eyes defocus, I initially see two overlapping images; but, these soon coalesce into a single 3-D image that can be quite startling.

If I *initiated* the eye movements involved, presumably by triggering some brain circuit into discharging in whatever way was necessary to achieve the desired effect, I can reasonably take credit for the overall result.

Taking credit for the result of actions I trigger certainly seems to be consistent with what Searle wrote about the self as a rational agent that can act despite the lack of neural conditions causally sufficient to determine the action taken; but, this amounts to saying that I can initiate (cause) actions that are uncaused by the brain; but, which are not random events.

However, if I claim that I supply the missing cause or that I initiate or trigger the action that brain conditions were not sufficient to cause, how do I avoid the conclusion that I have the property of being able to initiate the brain activity that carries out my intention? And how do I avoid the further conclusion that I am a substance distinct from my brain, making interactive substance dualism true?

On the other hand, if I try to avoid the charge of dualism by claiming to be causally reduced, I become an epiphenomenal I. I lack the ability to initiate action uncaused by the brain; so, nothing ever happens except that which can be explained in a deterministic way.

* * *

It seems that we have returned to the core of the Searlean Dilemma. Given that consciousness (*qua* subject) is not identical to its brain (or to some part thereof, or some brain state or brain process etc.), asserting the thesis of causal reducibility results in epiphenomenalism; but, the thesis of subject causation results in interactive substance dualism.

If the only other choice is epiphenomenalism, I choose interactive subject dualism. To be more precise, I choose the assumption – subject causation – that leads to interactive substance dualism rather than the assumption that leads to epiphenomenal dualism.

It might be argued that interactive substance dualism is not actually an option because it runs afoul of the causal closure principle which is wildly popular among physicalist philosophers and which Searle also supports.

> Surely the real physical world is 'causally closed' in the sense that nothing from outside the physical world can ever have any causal effects inside the physical world. (Searle, 2004, 193).

If the brain lacks conditions causally sufficient to determine the result; and, I, as a rational agent, somehow initiated action anyway, I caused a physical$_1$ effect. Assuming that I am the non-physical$_1$ cause of the physical$_1$ activity I initiate, I am the non-physical$_1$ cause of a physical$_1$ effect; and, *that* is precisely what the causal closure principle is an attempt to prohibit.

It is time to examine the causal closure principle.

### §4.5.3.3 Throwing Down the Gauntlet

The causal closure principle popularized by Jaegwon Kim is something like this:

> **[CCP] If a physical event has a cause at t, then it has a sufficient physical cause at t.**

It should be noted that Kim states his principle without the explicit use of "sufficient" that I've interpolated; but, he applies his principle as if it were present. When considering the possibility that a physical cause and a mental cause are jointly necessary but individually insufficient to produce an effect, Kim writes, "This seems like an absurd thing to say, and in any case it violates the causal closure principle in that it regards the mental event as a necessary constituent of a full cause of a physical event" (Kim, 1989, p. 44).

As explained by Kim, "According to this principle, physics is causally and explanatorily *self-sufficient*: there is no need to go outside the physical domain to find a cause, or a causal explanation, of a physical event" (Kim, 2005, p. 31).

The problem for philosophers of consciousness is that adding a free will postulate to an otherwise innocuous philosophy appears to yield dualism. Attempts to classify a resulting philosophy as a non-reductive materialism rather than a form of dualism run up against counterarguments relying on the principle of causal closure.

> I am supposing that a nonreductive physicalist is a mental realist, and that to be a mental realist, your mental properties must be causal properties – properties in virtue of which an event enters into causal relations it would otherwise not have entered into. (Kim, 1989, p. 43)

Kim then goes on to argue that nonreductive materialism so construed is in conflict with the causal closure principle; but, I won't critique that argument. I'm willing to assume that, if the causal closure principle is true, non-reductive materialism runs afoul of it.

I reject the causal closure principle itself, for the reason that it is anti-scientific.

However popular it is among physicalist philosophers, the causal closure of the physical is a principle of dubious provenance. At best, it is an extraneous

ingredient that physicalist philosophers add to the scientific account of the world. At worst, it is also a principle that physicists regularly contradict. In a chapter aptly titled “Physicalism Versus Quantum Mechanics”, physicist Henry P. Stapp wrote:

> An examination of the structure of quantum mechanics reveals that the theory has both a logical place for, and a logical need for, choices that are made in practice by the human actor/observers, but that are not determined by the quantum physical state of the entire world, or by any part of it. Bohr calls this choice “the free choice of experimental arrangement for which the quantum mechanical formalism offers the appropriate latitude.”
>
> The fact that this choice made by the human observer/agent is not determined by the physical state of the universe means that *the principle of the causal closure of the physical domain is not maintained in contemporary basic physical theory*. It means also that Kim’s formulation of *mind-body supervenience is not entailed by contemporary physical theory*. (Stapp, 2009, p. 248)

Stapp goes on to chide Kim for trying “to squash the notion that the difficulties with physicalism can be avoided by accepting some form of dualism” but only considering the form of dualism “advanced by Descartes during the seventeenth century instead of in the form employed in contemporary science” (Stapp, 2009, 249).

Unable to find an online reply by Kim to Stapp (2009), I emailed Kim to ask if he had published a reply to Stapp and to ask about the provenance of the causal closure principle. Although he replied to my email (Kim, 2015), he claimed to be unaware of Stapp's paper and did not identify his source for the causal closure principle. I sent Kim the link to the paper on Stapp's website; but, he never sent a reply to the challenge of “Physicalism Versus Quantum Mechanics”.[50]

### §4.5.3.4 The Causal Closure Principle of Physicalism

In an historical review of the causal closure principle, David Papineau (2001) noted that it is a key premise in a family of causal arguments for physicalism, “the doctrine that everything with causal powers is physical”.

Papineau argues that the persuasiveness of the causal arguments is so great that it accounts for the rise of physicalism. I concede that the causal closure principle is very popular among physicalistic philosophers; and, that it may well account for the widespread adherence to physicalistic philosophies of consciousness; but, in my view, its epistemological status is questionable.

The causal closure principle is not an empirical finding of physics. Papineau recalls being asked to state the grounds for believing this principle.

> However, when they then asked me, not unreasonably, to show them where the completeness of physics is written down in the physics textbooks, I found myself in some embarrassment. Once I was forced to defend it, I realized that the completeness of physics is by no means self-evident. (Papineau, 2001)

---

50 Surely, physicalist philosophers must overcome the discrepancy between what they say and what physicists say about the physical universe; otherwise, not to put too fine a point on this, it begins to look like the sorcerers' apprentices are running amok.

Another philosopher with a similar perspective is Augustin Vicente who concedes that “as far as I can see, there is no direct entailment from any physical law, or set of physical laws, to the CCP [Causal Closure Principle].” (2006, 157). Both Vicente and Papineau (before recanting) argued on largely inductive grounds, that we have good reasons for believing that the causal closure principle is true.

But, an inductive argument for a conclusion that scientists contradict is certainly a dubious argument.

### §4.5.3.5 The Free Will Postulate of Physics

> It is usually tacitly assumed that experimenters have sufficient free will to choose the settings of their apparatus in a way that is not determined by past history. We make this assumption explicit precisely because our theorem deduces from it the more surprising fact that the particles’ responses are also not determined by past history. (Conway & Kochen, 2006, 3-4)

Mathematicians John Horton Conway and Simon Kochen have proposed two theorems, the Free Will Theorem and the Stronger Free Will Theorem, in which they formalize the tacit assumption of physicists that they have the choice that Bohr indicated they have. These theorems show that, given the Free Will assumption and other assumptions, it follows that a particle's response to a measurement is unpredictable even in principle.

> Although, as we show in [Conway & Kochen, 2006], determinism may formally be shown to be consistent, there is no longer any evidence that supports it, in view of the fact that classical physics has been superseded by quantum mechanics, a non-deterministic theory. The import of the free will theorem is that it is not only current quantum theory, but the world itself that is non-deterministic, so that no future theory can return us to a clockwork universe. (Conway & Kochen, 2009, 230)

Once the formerly tacit assumption became explicit, it became available for further theorizing – with interesting results.

> Here we ask the more general question of whether any improved predictions can be achieved by any extension of quantum theory. Under the assumption that measurements can be chosen freely, we answer this question in the negative: no extension of quantum theory can give more information about the outcomes of future measurements than quantum theory itself. (Colbeck and Renner, 2011, 1)

An experimental test of this theorem was able to rule out present and future theories that made predictions significantly better than quantum theory. (Stuart et al, 2012, 3-4)

In a later paper, Colbeck & Renner defend the realist (objective or ψ-ontic) as opposed to the instrumentalist (subjective or ψ-epistemic) interpretation of the wave function. Their argument is that a quantum system's wave function is uniquely determined by its underlying physical state; and, therefore, it may be considered an objective property of the system (ie. an element of reality) rather than merely a representation of our incomplete knowledge about the system.

> ... the quantum wave function can be taken to be an element of reality of a system based on two assumptions, the correctness of quantum theory and the freedom of choice for measurement settings.

> The correctness of quantum theory is a natural assumption given that we are asking whether the quantum wave function is an element of reality of a system. Furthermore, a free choice assumption is necessary to show that the answer is yes. ... This shows that the wave function would admit a subjective interpretation if the free choice assumption was dropped. (Colbeck & Renner, 2012, 3-4)

This result bears on the question posed by Einstein, Podolsky and Rosen as to whether quantum theory can be considered complete. EPR argued that there were elements of reality that quantum theory could not discern, resulting in an incomplete theory. Bohr's reply was there there were no elements of reality that quantum theory was unable to discern. Given the assumption of freedom of choice for measurement settings, it would seem that Bohr was correct.

There are other theorems that conclude that the wave function is a real physical state; so, the Colbeck-Renner theorems can be viewed as an independent line of evidence for the same conclusion. As such it joins other theorems and experimental evidence in support of what Herbut (2014) calls the *ontic breakthrough* in quantum mechanics: increased support for the view that the wave function represents a state of reality rather than merely a state of knowledge.

The reality status of the wave function is a tremendously important issue for the science and philosophy of consciousness because, if the wave function is a physical entity, the collapse of the wave function is a physical event. That would raise a question to which there is no easy answer: *What causes wave function collapse?* For reviews of the status of the debate between advocates of ψ-ontic and ψ-epistemic views concerning the wavefunction, see Herbut (2014), Leifer (2014) and Dorato & Laudisa (2014).

One may also take the Colbeck-Renner theorem to show that the free choice assumption does not lead to a result that contradicts ψ-ontic theories; but, it also sets up a way to test the assumption of freedom of choice as to measurement settings. If a ψ-epistemic theory was shown to be correct, it would imply that there was no such freedom.

At the moment however, the question concerns the status of the free will postulate in physics; and, one naturally wonders just what this means to Colbeck and Renner and others who use it in their theorems.

The authors specify that “a choice A is free if it is uncorrelated with any other variables, except those that lie in the future of A in the chronological structure”. (Colbeck & Renner, 2013a, 3)

Clearly, if a choice is the causal result of a pre-existing brain condition, it would not meet that criterion; so, it is reasonable to infer that a free choice for these physicists is one that is *uncaused* by brain conditions – the traditional way of describing a “libertarian” free will.

A determined philosopher might be moved to say that a failure of determinism only establishes that indeterminism is the case; and, that further effort would be required to rule out randomness before a theory of intentional actions can be accepted.

> It may well be true that classically stochastic processes such as tossing a (true) coin do not help in explaining free will, but, … adding randomness also does not explain the quantum mechanical effects described in our theorem.
>
> ...
>
> In the present state of knowledge, it is certainly beyond our capabilities to understand the connection between the free decisions of particles and humans, but the free will of neither of these is accounted for by mere randomness. (Conway & Kochen, 2009, 230)

Stapp has also elaborated on what constitutes a free choice. Freely chosen actions are “not determined, via any known law, by the physically described state of the universe” and “seem to us to be freely chosen by our mental processes”. (Stapp, 2008, 25)

### §4.5.3.6 The Convergence of Physics and Philosophy

Suppose we reject (as I do) both Kim's causal closure principle and Searle's causal reduction principle as anti-scientific deadwood, extraneous ingredients not founded in physics. What then? Remarkably, the result of deleting both principles from Searle's theory of voluntary action turns out to be precisely what physicists need to explain quantum mechanics. Stapp explains ...

> The essential point here is that Searle escapes reduction of the mental to the physical by introducing, in addition to the physically described realities not simply conscious events, per se, but thinking agents. This is precisely the answer given by quantum triality. Quantum mechanics gives not a dualism of ‘physical things’ and ‘mental things’, but rather a triality consisting of: (1), the physically described aspects of reality; (2), conscious agents that first choose a physical probing actions, then initiate it, and finally register the response to the chosen action; and (3), a ‘nature’ that determines these responses. The second aspect consists of precisely the “entities” that Searle demands. (Stapp, 2010, 8)

Now, let us recall that Searle concluded his inquiry into the conditions that would make voluntary actions possible by saying that, if we reject the epiphenomenal option, we would end up with three inter-related mysteries: free will, consciousness and quantum mechanics. Searle went on to say that “in order to solve the first we have to solve the second and invoke one of the most mysterious aspects of the third [quantum indeterminism] to solve the first two”. (Searle, 2006, 78)

This is a remarkable convergence of physics and philosophy!

There is a downside, however. If I am a conscious agent able to make choices and initiate actions undetermined by the prior state of my brain, I have the property of being able to make such choices and the property of being able to initiate actions that implement my choices. From the philosopher's perspective, any such theory makes me a property bearer; and, therefore, a substance.

Consequently, from the philosopher's perspective, a physical theory that incorporates a free will postulate constitutes interactive substance dualism.[51]

---

51 Interactive substance dualism is generally thought to be incompatible with physicalism; but, in my view, tests for physicalism should be independent of tests for dualism. Any philosophy

However, without assumptions not yet made and/or conclusions not yet drawn, it would not constitute Cartesian-style interactive substance dualism, which I will take to require two self-existent substances each of which may be considered a kind of “stuff”.

Instead, the result of cleaning up Searle's theory of the rational agent yields a weaker form of interactive substance dualism, a minimalistic substance dualism, similar to the version of dualistic emergence advocated by Martine Nida-Rümelin or the non-Cartesian substance dualism advocated by E. J. Lowe.

### §4.5.3.6.1 Dualistic Emergence

According to Nida-Rümelin, at some point in the life of a human a conscious being comes into existence.

> The emergentist believes that this change occurs as a result of physical conditions satisfied by the biological system. A certain arrangement of matter leads with nomological necessity to the existence of conscious individuals with qualitatively new properties. (Nida-Rümelin, 2006, 2)

For Nida-Rümelin, a conscious being is a subject of experience. She uses the terms interchangeably; so, we have something very close to Searle's position. Searle argues against the zombie argument for dualism by saying that

> This argument is sometimes put in the form as an imagined parable about the creation of the world. Imagine God creating the world. First he has to create all the physical particles. Then he has to add the laws that determine the behavior of the physical particles. And finally, after He has done all that, He still has to add consciousness. He might have to add some more laws to get consciousness, but consciousness is something in addition to the physical particles and physical behavior. ...
>
> On my view, given the constitution of reality, consciousness has to follow in the same way that any other biological property, such as mitosis, meiosis, photosynthesis, digestion, lactation, or the secretion of bile, follows. (Searle, 2007a, 177)

Consciousness may follow for Searle; but, it is not ontologically reducible because it is not identical to the system of interacting brain processes from which it emerges.

For Nida-Rümelin, the conscious being that comes into existence “is not identical to the system that gives rise to its occurrence”. (2006, 3)

Where Searle would say that the subject of experience causes his arm to rise, Nida-Rümelin would say that the subject “is a causal origin of changes in the brain that initiate and that uphold a movement when the subject does something involving a bodily movement”. (2006, 9)

Searle says that certain actions occur despite the lack of brain conditions causally sufficient to determine the result. Nida-Rümelin says that “the subject can causally influence physical events happening in its own brain”. (2006, 11)

---

consistent with some interpretation of quantum mechanics has as much right to be called physicalism as any other such theory. If a theory describes or explains something about brain/consciousness relations by reference to a pair of somethings, it may be considered a form of dualism. If a theory meets both tests, it may be considered a dualistic physicalism.

These similarities between Searle and Nida-Rümelin naturally suggest that their viewpoints should be similarly classified; but, while Searle denies being a dualist, Nida-Rümelin argues that her viewpoint constitutes a form of substance dualism.

One might challenge Nida-Rümelin's claim to be a substance dualist by challenging the conception of substance as property bearer. Tim Crane (2003) argues that being a property bearer is not sufficient to be a substance. All *particulars* have properties but not all particulars are substances; events (occurrents) are his example of a property bearing particular which is not a substance (continuant). Assuming arguendo that a substance must be a continuant, Nida-Rümelin (2006, 5) meets that condition. “What a subject of experience is can best be positively characterized by saying that it is capable of having consciousness properties and by describing the special ontological status of its identity across time and of its identity across possible worlds.”

One might also challenge Nida-Rümelin's claim to be a substance dualist on the ground that, for her, a conscious being is a dependent substance rather than an independent (self-existent) substance as it is for Descartes. She admits that

> If the term 'substance' is reserved to entities that do not depend for their existence on the existence of any other entity, then the view proposed cannot be classified as a version of substance dualism. (Nida-Rümelin, 2010, 191)

There is no logical reason that prevents a philosopher from weakening the concept of 'substance' to allow for dependent substances. Once that step is taken, one ends up with a form of dualism that is stronger than an attempt (by Chalmers, for example) to account for experiential phenomena by postulating a second set of properties for a single property bearer (the brain or matter generally).

In any case, Nida-Rümelin *is* a substance dualist according to my definition because she postulates a second property bearer, the subject of experience, to instantiate the property of being able to initiate/cause brain events and subsequent bodily actions.

To avoid Cartesian-style interactive substance dualism, Nida-Rümelin explicitly rejects several of the claims usually associated with Descartes. By claiming that the subject of experience comes into existence when bodily conditions support its existence and ceases to exist when the body dies, she denies the identification of the subject with the soul as traditionally conceived. She also denies that there is any “thin” non-material stuff out of which the subject is made.

In contrast to Nida-Rümelin, Searle denies being a dualist; but, the comparison with Nida-Rümelin suggests that postulating an irreducible self to defend the free will postulate brings with it the challenge of avoiding interactive substance dualism, the ultimate philosophical catastrophe for those who deny being dualists.

### §4.5.3.6.2 Non-Cartesian Substance Dualism

The late E. J. Lowe developed what he called NCSD, Non-Cartesian Substance Dualism, based on two essential principles: the non-identity of the I, a person or self, and its brain; and, the causal efficacy of the I. According to Lowe, NCSD …

> … regards persons as substances in their own right, in the sense of 'substance' in which this denotes a persisting entity and bearer of properties which does not depend for its identity on anything other than itself. (Lowe, 2006, 5)

Like Nida-Rümelin, Lowe rejects the “stuff” interpretation of substance dualism; although, perhaps, not as firmly as Nida-Rümelin. Interestingly enough (but irrelevant for present purposes), Lowe denies that Descartes held a “stuff” view of substance dualism.

For Lowe, the self is still dependent for its existence on its brain; so, classifying Lowe as a substance dualist is still a matter of defining “substance” to mean “property bearer”.

Lowe's view of intentional action is similar to Searle's.

### §4.5.3.6.3 Reply to Nida-Rümelin and Lowe

So, while I can agree with Nida-Rümelin and Lowe that there are good reasons for me to assert “I am this subject of experience”, I don't thereby acquire knowledge of my causal origin, how I came to be. I *could* make an additional assumption, as Nida-Rümelin does, and assume that I came into being when my body created the conditions necessary for my existence; but, I don't; and, neither do I make the opposite assumption, that I existed prior to the time that my body first created the conditions necessary for the manifestation of my existence. Similarly, I *could* assume that I cease to be when my body ceases to support my existence; but, I don't; and, neither do I make the opposite assumption that I continue to be after my body ceases to support my existence.

At least at this point in the inquiry, I have no reason to make those assumptions. All I know is that I am a subject of experience *now*.

On the other hand, I disagree with Nida-Rümelin concerning the supervenience principle. Supervenience is a dependence relation; and, in Nida-Rümelin's view, consciousness is dependent on the brain for its existence. Unfortunately, it is not at all clear that such a theory of consciousness is consistent with a theory of physics that assumes a free will postulate.

Any number of difficult questions might be raised at this point. Would a consciousness generated by the brain be subject to the Schrodinger equation? If so, how could it collapse the wave function instead of becoming entangled with it? All interpretations of quantum mechanics in which consciousness causes collapse upon measurement would have to be false before Nida-Rümelin's view could be true. How does the brain generate a consciousness with the property of being able to act independently of the brain – to trigger brain events uncaused by the brain?

The alternative, assuming that consciousness (qua subject/agent) is not

dependent on the brain for its existence, poses equally difficult challenges. For one thing, it would be impossible to escape the conclusion that the ontology of physics has been expanded. The result may still be considered a form of physicalism; but, it would be a dualistic form of physicalism.

#### §4.5.3.6.4 Physics and Subject/Body Dualism

Stapp disclaims substance dualism but he contests the popular view of Cartesianism – substance as ectoplasmic "stuff" – and the idea of substance as ordinary stuff as well.

> It is worth noting that the physically described aspect of the theory has lost its character of being a "substance", both in the philosophical sense that it is no longer self-sufficient, being intrinsically and dynamically linked to the mental, and also in the colloquial sense of no longer being material. It is stripped of materiality by its character of being merely a potentiality or possibility for a future event. This shift in its basic character renders the physical aspect somewhat idea-like, even though it is conceived to represent objectively real tendencies. (Stapp, 2009, 248)

He affirms a form of dualism where a substance is merely a *carrier of essences*.

> This solution is in line with Descartes' idea of two 'substances', that can interact in our brains, provided 'substance' means merely a carrier of 'essences'. The essence of the inhabitants of res cogitans is 'felt experience'. They are thoughts, ideas, and feelings: the realities that hang together to form our streams of conscious experiences. But the essence of the inhabitants of res extensa is not at all that of the sort of persisting stuff that classical physicists imagined the physical world to be made of. (Stapp, 2007, 167)

If *carrier of essences* means more or less what *property bearer* means, then Stapp would be an interactive substance dualist.

## *§4.5.4 The Epiphenomenal Option*

What I'm calling the *Searlean dilemma* arises after one affirms the existence of consciousness (qua subject) and denies the identity of the experiencing subject and its brain. To have a complete theory of consciousness, one must address the question of subject initiated causation – mental causation or intentional action. But, at this point in the epistemological journey, our choices are limited to epiphenomenal dualism and interactive dualism, each of which is considered unpalatable by many philosophers.

By his doctrine of causal reducibility, Searle appears to avoid interactive dualism at the expense of embracing epiphenomenalism. If I am correct about the effect of assuming causal reducibility, Searle will not be able to deny being a dualist on the grounds that dualism leads to epiphenomenalism but his views do not. But that conclusion about Searle, even if fully justified, doesn't tell us anything about the relative merits and demerits of interactive and epiphenomenal dualism.

Interactive dualism is certainly consistent with the common sense view of life. Furthermore, it has the virtue of being consistent the views of physicists who assume a free will postulate.

Nevertheless, some philosophers seem to prefer the other side of the Searlean

dilemma: being an epiphenomenon. I'll now consider the case for and against epiphenomenal dualism.

### §4.5.4.1 Terminology

I'll use the unqualified terms, “interactionism” and “epiphenomenalism” this way:

> **Interactionism** is the position holding that something not physical$_1$ – something experiential (or mental or spiritual or whatever) – has an effect on something physical$_1$ (and vice versa).
>
> **Epiphenomenalism** is the position holding that nothing non-physical$_1$ ever has any effect on anything physical$_1$; sometimes called general epiphenomenalism.

There is also a position of limited epiphenomenalism that has a number of advocates.

> **Qualia Epiphenomenalism** is the position holding that experiential phenomena (qualia) have no causal effects on anything physical$_1$.

There does not seem to be a term in the literature for the claim that qualia are causally efficacious, so I'll propose the following:

> **Qualia Interactionism** is the position holding that experiential phenomena have causal effects (directly or indirectly) on something physical$_1$.

Qualia epiphenomenalists may but need not hold that something besides experiential phenomena that is also not physical$_1$ is involved in events in which something mental has a causal effect on something physical$_1$.

One obvious question concerns the causal role (if any) of the subject of experience (if any). To put this in terms of the Knowledge Argument, let us suppose that, upon her release, Mary is presented with a tomato and reports “I am now experiencing *tomato red!*”. After staring at her tomato for sometime, she looks away and experiences an afterimage – her first – of the color complement type. While that experience is ongoing, Mary reports “I am now experiencing *otamot green!*”.

How do we account for these reports? Are they examples of brain initiated causation, quale initiated causation, subject initiated causation or something else? Did Mary (qua subject of experience) have any choice in the manner of communicating her report? After tweeting her initial reports to the Friends of Mary network, let's assume that Mary issues a third report, this time by email, “I chose to email this report instead of tweeting it to you”.

Did Mary (qua experiencing subject) initiate the causal events that culminated in each of her reports? Did a quale she experienced make her do it? Did her brain make her do it?

To allow discussion of these issues, two more terms may be helpful. The first is the position I've previously defined.

> **Subject Causation** is the position that the experiencing subject exists, is not identical to its brain and is, at least some of the time, the agent of causation having a physical$_1$ effect.

Someone may deny subject causation, deny that the experiencing subject is ever the agent or initiator of causation, by affirming subject epiphenomenalism.

> **Subject Epiphenomenalism** is the position holding that the experiencing subject exists, is not identical to its brain and is never the agent of causation having a physical$_1$ effect.

### §4.5.4.2 A Tale of Two Claims

In preparing his readers for the structure of "Epiphenomenal Qualia", the paper in which he introduced his Knowledge Argument, Jackson noted insightfully that "The major factor in stopping people from admitting qualia is the belief that they would have to be given a causal role with respect to the physical world and especially the brain" (1982, 128), citing Dennett (1978) in a footnote; presumably, for the following passage:

> Suppose, with the dualists, that there are non-physical effects (or accompaniments) of brain events. Then either the occurrence of these effects has itself *no effect whatsoever* on subsequent events in the brain (and hence behavior) of the person (epiphenomenalism), or it does (interactionistic or Cartesian dualism). In the former case the postulation of the non-physical effects is utterly idle, for *ex hypothesi* were the effects to cease to occur (other things remaining the same), people would go right on making the same sorts of introspective claims, avowing their pains, and taking as much aspirin as ever. … In the latter case of interactionistic dualism, since the occurrence of non-physical events (events having temporal location and presumably particular person dependency but lacking spatial location and mass-energy) would be required to trigger unproblematically physical events in the brain, the conservation laws of physics would be violated. Either way, one pays an exorbitant price for dualism. (Dennett, 1978, 252)

Once we admit that brain activity is accompanied by experiential phenomena not identical to any of that brain activity, we have some form of dualism. We must choose between epiphenomenal dualism and interactive dualism.[52]

Clearly, since Jackson presented the KA as an argument for the existence of non-physical qualia and defended epiphenomenalism against some common objections, we may infer that Jackson quietly chose epiphenomenal dualism over interactive dualism; but, sadly, he did not present an affirmative argument showing that qualia are epiphenomenal. Instead, he asks, "Is there any really good reason for refusing to countenance the idea that qualia are causally impotent with respect to the physical world? I will argue for the answer no ..." (Jackson, 1982, 133) and proceeds to offer his defenses to common objections to epiphenomenalism.

There are any number of replies to Jackson's KA; but, I will focus on those that

---

52 However, Dennett's suggestion that any form of interactionistic dualism is equivalent to Cartesian dualism is quite unhelpful. A claim that the color of her first afterimage makes a causal contribution to Mary's report, "I am now experiencing otamot green" is not equivalent to a claim that humans are (or have) souls which survive the deaths of their bodies.

make up what, following Nagasawa (2010), I will call the Inconsistency Objection – the allegation that the claim of knowledge is in conflict with the claim of epiphenomenalism. The claim of knowledge is the inference drawn from the knowledge argument itself (i.e. the argument in part I of “Epiphenomenal Qualia” as restated in Jackson (1986) to the effect that Mary acquires knowledge from her encounter with color qualia). The claim of epiphenomenalism is the claim that the experiential color phenomena Mary experiences following her release are epiphenomenal.

Jackson calls the Inconsistency Objection the 'There Must be a Reply' reply because …

> This reply does not tell us what is wrong with the knowledge argument. It seeks to show that there must be something wrong with it somewhere. What we know about the way the world works tells us that Mary *cannot* acquire knowledge of how things are that outruns the physical story she knew beforehand when she leaves the room – despite the fact that it certainly seems that she does! It seems to us to be the most powerful reply to the knowledge argument. It is what makes the phenomenal side of psychology such a hard problem. We have a good argument – the knowledge argument – that the physicalist picture is inadequate; yet we have a good argument – from causal considerations – that it must be adequate! (Braddon-Mitchell and Jackson, 2007, 142).

I can't speak for all dualists; but, I doubt that it seems to Mary that she learns something new about how physical$_1$ things are just by looking at a tomato; and, for the sake of the argument, I am willing to assume that she does not. In my view, Mary learns something new about how experiencing is. The question becomes whether *this* knowledge escapes or outruns the physical story; but, that question can not be answered without an account of what constitutes escaping or outrunning the story as told by this, that or some other variety of physicalism.

Above all, we also need an account of what is physical. If physicists are required to postulate an immaterial consciousness to explain wavefunction collapse or to support the Free Will Postulate; then, arguably, it may be considered physical in some sense. But then not even interactive substance dualism will escape the physical story.

Using a maximally restrictive definition of “physical”, where only what is physical$_1$ is physical, one only needs to escape the story as told by eliminative materialists, type-Z materialists and identity theory physicalists to escape the physicalist story.

Using a definition somewhat less restrictive, one might argue that experiential phenomena not identical to any physical$_1$ phenomenon may still be physical if they are explainable as causal effects of physical$_1$ phenomena. Now we would have a more ambiguous outcome. Knowledge by acquaintance of experiential phenomena would not escape the story told by promissory materialists (and their counterparts among physicalists). We would have to wait until we have a completed physics to discover whether what Mary learns by her first experiences with color qualia escapes the physicalist story.

On the other hand, with such a definition of “physical” the conversation might just shift to a discussion of whether what Mary learns is encompassed by the story told by present day physical science. In my view the answer to that

question will by “No” for as long as promissory materialists are still hoping for deliverance.

I will set aside these and any other concerns about the criteria for evaluating claims of success or failure at explaining qualia, refuting physicalism, etc. to focus on the structure of the Inconsistency Objection.[53]

> **[IO-1] If there is a sound argument for epiphenomenalism, Mary cannot acquire new knowledge of experiential phenomena upon her release.**
>
> **[IO-2] If the knowledge argument is sound, Mary acquires new knowledge of experiential phenomena upon her release.**
>
> **[IO-3] (Therefore) If the knowledge argument is sound, there is no sound argument for epiphenomenalism; and, if there is a sound argument for epiphenomenalism, the knowledge argument is unsound.**

We have arrived at what I will call *Jackson's Dilemma*. one must reject either the knowledge argument or the argument for epiphenomenalism.

### §4.5.4.2.1 Targeting the Claim of Knowledge

One version of the Inconsistency Objection advanced independently by Stjernberg (1996) and Watkins (1989) assumes that there must be a causal link between an experiential phenomenon and our knowledge of it. This creates an problem for anyone claiming to know that experiential phenomena are epiphenomenal. As Michael Watkins explains,

> If qualia are not causally efficacious, then my beliefs and memories would be just as they are whether there were qualia or not. Beliefs about qualia cannot be justified on the basis of qualitative experiences since those experiences do not cause those beliefs. … The only evidence we have of qualia is our direct experience of them. On Jackson's story, however, we are told that our beliefs concerning qualia are actually caused by brain states and would be the same whether the qualia exist or not.
>
> Mary, the heroine of Jackson's knowledge argument against physicalism, gains no new knowledge when she leaves her black and white room, only unjustified beliefs. …

Watkins seems to be trying to entice Jackson back into the physicalist fold, for he continues:

> In order to allow for our knowledge of qualia Jackson must either retreat to interaction dualism, pay homage to some mysterious parallelism, or else join the ranks of physicalism in the hope that the qualia problem will find a happy resolution. (Watkins, 1989, 160)

As is well known, Jackson has indeed returned to the ranks of physicalism; although, I don't know whether this is the argument that convinced him that

---

53 This version of the Inconsistency Objection is (very loosely) based on the version offered by Nagasawa (2010). While it shrinks from five propositions to three, the main difference is that my version speaks about the argument for epiphenomenalism being sound whereas Nagasawa's version speaks about epiphenomenalism being true.

qualia epiphenomenalism is untenable.

> If certain mental states have qualia in the sense of properties that fall outside the physicalists' picture, these qualia must be epiphenomenal. This follows from the considerations of causal closure discussed in chapter 1 … But then beliefs and memories cannot be regarded as responses to the existence of qualia. The exposition by qualia freaks of the knowledge argument can be in no sense the *outcome* of the instantiation of qualia." (Braddon-Mitchell and Jackson, 2007, 141)

Those who hold that Mary gains knowledge rather than an unjustified belief from her first encounter with color qualia need a reply to this version of the Inconsistency Objection.

#### §4.5.4.2.1.1 Acquaintance Is Not Causal

One strategy is to deny that the causal theory of knowledge applies to experiential phenomena. Chalmers explains.

> In response to the argument from the causal theory of knowledge, we note that there is independent reason to believe that the causal theory is inappropriate to explicate our knowledge of experience: our knowledge of experience is grounded in a more immediate relation. … the justification of my beliefs about experience involves more than the mechanisms by which the beliefs are formed: it crucially involves experiences themselves. (Chalmers, 1996, 198)

In a later work, Chalmers develops this view further, concluding that there is a special epistemic relation holding between the experiencing subject and "the phenomenal properties instantiated in our experience" and that "we might call this relation *acquaintance*" (2003, 248).

Significantly, William S. Robinson a philosopher who has defended qualia epiphenomenalism for many years, acknowledges that Chalmers' position avoids any conflict with the causal theory of knowledge.

> In supplying non-causal relations to support the claim to knowledge of experiences, this view disconnects the knowledge question from the question of how things stand causally, and thus avoids the self-stultification argument. (Robinson, 2015)[54]

An acquaintance based reply may also defeat other versions of the Inconsistency Objection. For example, consider the version offered by Neil Campbell (2003). Campbell argues that there is tension in Jackson's KA.

> The source of the tension is that his argument for the non-physical character of qualia is plausible only on the assumption that they have causal efficacy, while his argument for the epiphenomenal character of qualia is plausible only on the assumption that they are non-physical. Since these two arguments cannot be combined coherently, the most Jackson's argument can establish is that qualia are non-physical. (Campbell, 2003. 261).

Not everyone wants Jackson's argument to establish more than that qualia are non-physical; so, this fact would seem to be a virtue rather than a drawback to

---

54 It should be noted that Robinson does not endorse the Acquaintance Reply to the Inconsistency Objection. Indeed, as will be discussed below, he seems to deny the possibility of knowing by acquaintance.

anyone who rejects epiphenomenalism.

> Jackson's conclusion that qualia are non-physical depends on our readiness to accept the claim that Mary learns something new when she leaves her black-and-white environment, but there is no reason to suppose Mary learns anything new unless her qualia are causally efficacious. (Campbell, 2003. 262).

The flaw in Campbell's argument is that he doesn't consider the Acquaintance Reply. We have every reason to suppose that Mary learns something new when she becomes acquainted with color qualia; but, if acquaintance is not a causal process, it doesn't matter that *qualia* are not causally efficacious; provided, that the experiencing subject is causally effective.

#### §4.5.4.2.1.2 Limited Epiphenomenalism

Another strategy mentioned by both Stjernberg and Nagasawa involves limiting the claim of epiphenomenalism.

> Yet another way to meet the objection is to partly give up the idea that qualia are epiphenomenal, and say that the instantiation of qualia has causal effects on other mental states. Given what we may want to say about other mental phenomena such as remembering, it is desirable that qualia should have some causal effects on other mental states, at least. (How else could one remember what a qualia [*sic.*] was like? Why would one want to have that chocolate cake again?) This idea is still difficult to assess, since it appears to be *prima facie* plausible to hold that mental phenomena are connected, which here would mean that either all mental phenomena are epiphenomenal, or none are. Therefore epiphenomenal qualia lead to general epiphenomenalism, or else they are not really epiphenomenal. (Stjernberg, 1996, 10)

Interestingly, this approach was suggested by Jackson himself. While defending the claim that qualia are epiphenomenal, he left open the possibility that something else could be mental without being epiphenomenal.

> I will say nothing about two views associated with the classical epiphenomenalist position. The first is that mental states are inefficacious with respect to the physical world. All I will be concerned to defend is that it is possible to hold that certain properties of certain mental states, namely those I've called qualia, are such that their possession or absence makes no difference to the physical world. The second is that the mental is totally causally inefficacious. For all I will say it may be that you have to hold that the instantiation of qualia makes a difference to other mental states though not to anything physical. Indeed general considerations to do with how you could come to be aware of the instantiation of qualia suggest such a position. (Jackson, 1982, 133)

Hans Muller (2008) presents a version of the Inconsistency Objection targeting the attempt to limit one's epiphenomenalism to qualia epiphenomenalism.

> If Jackson had never had the experience of feeling pain (i.e., having that quale) he would not have written his now famous article in which he tried to convince the rest of us that qualia are both real and causally impotent with respect to the physical world. ... [O]ne seemingly cannot say, 'I have qualia and they are causally inert' without falsifying that very claim via the act of asserting it. (Muller, 2008, 88)

In a reply to Muller, Dan Cavedon-Taylor (2009) argues that Muller did not catch Jackson in a contradiction because

> Muller provides no reason for thinking that it is Jackson's qualia itself which causally interacted with the physical world and which caused him to write "Epiphenomenal Qualia" instead of, say, Jackson's belief in qualia. It seems to me that Jackson can respond to Muller's argument by claiming that his qualia caused him to have the belief that he possesses qualia and that it is this later, belief, state that is the direct cause of his writing the paper in question. On this explanation, the epiphenomenal status of qualia is preserved insofar as such states fail to directly causally interact with the physical world; rather, Jackson's belief in the existence of qualia directly causally interacts with the physical world by bringing about, perhaps with the help of other attitudinal states, an assertion of qualia's existence. (Cavedon-Taylor, 2009, 105-106)

In a subsequent reply to Cavedon-Taylor, Muller points out that the qualia epiphenomenalist is still caught in a nasty dilemma.

> One option is to stand firm on the thesis that qualia have neither direct nor indirect causal power with respect to the physical realm. But as he have seen, this would mean that asserting the existence of qualia falsifies the theory. The second option is to ... accept the idea that qualia can cause mental states that in turn cause behavior and other physical changes. But as we have seen, this runs afoul of Jackson's claim that things with specifiable functional roles are physical things. (Muller, 2009, 112)

Dwayne Moore argues for the same dilemma.

> Assuming that qualitative properties causally influence the instantiation of mental intentional properties (beliefs/memories), a key question that arises is whether these intentional properties are reducible to functional/physical properties or not. (Moore, 2012b, 407)

If intentional properties are irreducibly mental, qualia would be able to causally contribute to beliefs and memories; but, those beliefs and memories would not be able to causally contribute to the behavior they are invoked to explain, an experiencing subjects' report about experiential phenomena.

On the other hand, if intentional properties are reducible to the physical, epiphenomenal qualia would be unable to affect them.

> If intentional properties are physical, as this solution presumes, then qualitative properties will not be able to causally contribute to the occurrence of these physical intentional properties. … And, if qualitative properties contribute nothing to the occurrence of beliefs/memories about qualitative properties, then those beliefs once again lack proper justification and cannot be counted as knowledge. (Moore, 2012b, 409)

This dilemma becomes more acute once we consider the role of the experiencing subject. Suppose Mary denies that the quale of phenomenal redness made her tweet her report instead of emailing it, claiming instead, "I, this experiencing subject, chose to tweet my report, 'I am now experiencing *tomato red*!'".

Now, a subject causationist would claim that the report was indirectly caused by the quale via its effect on the experiencing subject who then caused the issuance of the report.

*Could a qualia epiphenomenalist agree with such a position?*

There is no logical reason why a qualia epiphenomenalist can't agree with the subject causationist that the experiencing subject exists, is not identical to its

brain and is sometimes the agent of causation having a physical effect; but, I doubt that any philosopher would call that position epiphenomenal dualism rather than interactive dualism.

Perhaps the qualia epiphenomenalist would only agree that the experiencing subject exists and is not identical to its brain; denying that the subject is ever the agent of causation having a physical effect. This is the position I've defined as subject epiphenomenalism. For those philosophers defending a limited epiphenomenalism, the challenge is explaining how the combination of qualia epiphenomenalism plus subject epiphenomenalism does not constitute general epiphenomenalism.

Perhaps the qualia epiphenomenalist would retreat even further from the subject causationist position; denying subject/brain non-identity. But, if the subject is identical to its brain, it is a physical$_1$ object and qualia epiphenomenalist would have the choice of (1), denying that experiential phenomena have any effect on the experiencing brain; or, (2), explaining why any such effect does not constitute an interaction forbidden by epiphenomenalism.[55]

Finally, the qualia epiphenomenalist could deny the existence of the experiencing subject altogether; but, this approach increases the argumentative burden of the qualia epiphenomenalist. In addition to the burden of defending eliminativism as to the subject of experience, qualia epiphenomenalists would face the same choice confronting the qualia epiphenomenalist who asserts subject/brain identity.

#### §4.5.4.2.2 Targeting the Claim of Epiphenomenalism

Current arguments for epiphenomenalism presuppose the causal closure principle. As Nagasawa puts it, “the main motivation for holding epiphenomenalism is to preserve the causal closure of the physical” (2010, 41). Howard Robinson concurs. “We face a straight choice between accepting physical closure and accepting the knowledge argument” (2008, 224).

If so, then the dialectical situation in which we find ourselves points to the most decisive reply to the Inconsistency Objection, the Jugular Attack: deny causal closure, thereby undermining an assumption presupposed by arguments for epiphenomenalism. If there is no sound argument for epiphenomenalism there is one may deny it with impunity; thereby eliminating the appearance of inconsistency.[56]

As we've seen, the CCP is not only not a finding of physics, it's a “principle” that

---

55 This is a general problem for the limited epiphenomenalism approach: to the interactionist, it looks like interactive dualism. Even if the topic of the status of the subject is avoided, having a sequence of causal events beginning with a mental or experiential event such as the occurrence of a pain quale and ending with an undeniably physical event such as a report of experiencing pain is what interactionists call an interaction, experiential phenomena having an effect – directly or indirectly – on something physical$_1$.

56 If the CCP is false there is no sound argument for epiphenomenalism. We wouldn't be able to conclude that the knowledge argument is sound; but, we would be able to conclude that [IO-1] is only vacuously true.

physicists reject in the practice of physics. Nevertheless, arguments for epiphenomenalism invariably assume the causal closure of the physical domain.

> I think that for all we know epiphenomenalism may be true. I am not particularly fond of this doctrine. But I find the arguments leading to it very strong and the arguments trying to secure a causal role for the mind less than convincing. (Bieri, 1992, 283)

The arguments that prompt Bieri to conclude that epiphenomenalism is unavoidable are arguments based on causal closure and causal exclusion (which assumes causal closure).

William S. Robinson analyzed an argument "which purports to refute any dualism which includes the claim that we have non-inferential knowledge of our sensations" (1982, 524). Premise 3, "The causes of physical events are all physical", is a version of the causal closure principle, CCP.

In Robinson's later works, causal closure is not explicitly mentioned; but, it is implicitly assumed by alternate premises holding that behavior is completely determined. For example, "Our behavior – nonlinguistic and linguistic – is completely accounted for by the activity of our neuromuscular (and perhaps glial and microtubular) systems" (Robinson, 2004, 159). After setting aside the possibility of systematic overdetermination, Robinson draws the conclusion that phenomenal consciousness is causally ineffective. In Robinson (2006), there is a similar premise about behavior; but, the rejection of overdetermination is now explicitly taken as a premise.

Further examples of arguments for epiphenomenalism that rely on the CCP are easily found; for example, Seager (2006) and Gadenne (2006). In each case, the possibility of *underdetermination* by physical causes, is overlooked; presumably because it is implicitly ruled out by CCP or by a causal exclusion argument based on the CCP.

Epiphenomenalists also rely on the CCP when defending themselves against objections to their positions. For example, Volker Gadenne (2006) argues for qualia epiphenomenalism, QEP, based on causal closure and causal exclusion; defends QEP from various objections; and, steps back to reflect on what he has accomplished.

> The above considerations do not prove QEP. They rather should demonstrate that QEP is a reasonable possibility that cannot be easily rejected. QEP is not free from difficulties, but the decisive question is whether there are more convincing solutions. The main competing theories of mind are physicalism, which denies the nonphysical entities (including things and properties) exist, and interactionism, which is here conceived as interactionist property dualism, not as substance dualism. (Gadenne, 2006, 113)

After rejecting both reductive and nonreductive physicalism for various reasons, Gadenne says, "Interactionism, on the other hand, is incompatible with the principle of causal closure of the physical world. Epiphenomenalists have accepted causal closure.". (Gadenne, 2006, 113)

These uses of the CCP obviously beg the question against the interactionist and anyone else who denies the CCP or who rejects it as anti-scientific. Normally, a philosopher relying on a contentious assumption is expected to justify accepting

it; and, Gadenne offers this justification:

> Though I have tried to defend QEP, it seems to me that causal closure is not as evident as many contemporary philosophers of mind believe. It is often treated like a dogma. After all, causal closure can neither be established a priori nor is it dictated by physical science itself. It is mainly *methodological* considerations that suggest provisionally accepting causal closure. Roughly put, why should scientists postulate nonphysical Q-events as causes of the electrochemical activity of neurons, if they can well explain that activity as an effect of physical stimulation, or of electrochemical processes in other neurons? (Gadenne, 2006, 113-114)

The interactionist would probably agree that the CCP is held dogmatically by most physicalists and some dualists; meaning, that it is asserted and relied on but is not itself defended. Further, the interactionist would likely deny that there is any good reason to accept causal closure even provisionally and may even argue that it should be rejected altogether due to recent developments in physics discussed earlier.

Naturally, the epiphenomenal dualist may reply that the interactive dualist begs the question against epiphenomenal dualists by assuming a Free Will Postulate, FWP, such as that which many (but not all) physicists assume.

The interactive dualist could point out that deterministic theories of physics are not only completely deterministic in a way reminiscent of Laplacian determinism, they have a conspiratorial twist (Lewis, 2006). This state of affairs is popularly known as *super-determinism*; meaning, that determinating influences compel physicists to conduct experiments in such a way that they draw a *false* conclusion concerning the laws of physics.

This happens because, in testing Bell's Theorem, it is assumed that the experimenter has a free choice as to which experiment to conduct and when. All such experiments are looking at correlations between two sets of measurements in situations where quantum theory predicts one result and hidden variable theories predict another. To date, such experiments have produced results supporting orthodox quantum theory against hidden variable theories.

*But what if the game were rigged?*

Consider the possibility of counting cards while playing blackjack in a casino. Assume that the cards are shuffled honestly; so, we can expect a random sequence of cards to be dealt. Someone counting cards might notice occasional short run anomalies in which an above average number of low cards had already been dealt. That would mean that the remainder of the deck had a higher than average number of high cards left to play. Drawing high cards is generally good in blackjack; so, the card counter increases the size of his or her bets when the odds of getting a high card is higher than normal.

Card counting works (which is why casinos generally regard it as cheating). The MIT Blackjack Team, a group of students and ex-students, beat casinos all over the world by counting cards (Wikipedia, 2016-07-03).

According to super-deterministic theories of physics, determinating influences arrange for physicists to conduct their experiments only when statistical anomalies are about to occur. Experiments thus yield results from which faulty

conclusions are drawn about the nature of the universe.

Such an interpretation of quantum mechanics invites a number of objections. First, it undermines the validity of the entire scientific enterprise by running counter to the usual assumption that nature may be subtle but is not malicious.

Second, if epiphenomenal dualists rely on an interpretation of quantum mechanics which holds that nature deceives scientists, they would be hard pressed to explain why nature does not deceive epiphenomenal dualists instead.

Third, if epiphenomenal dualists can explain why they are not deceived, they would recreate the special people problem. Why would nature deceive most physicists but illuminate some philosophers concerning the way nature operates? Would armchair cogitation suddenly become more accurate than double slit experiments?

Reliance on super-deterministic interpretations of quantum mechanics makes epiphenomenal dualism even more counter-intuitive than it already is. On the other hand, without super-determinism, there is no basis in physics for rejecting the Free Will Postulate; and, epiphenomenal dualists must find an alternate justification for assuming causal closure. Without causal closure there is no argument for epiphenomenalism.

### §4.5.4.3 Epiphenomenal Determinism

We will have to wait to see how epiphenomenal dualists adapt to recent developments in physics. Meanwhile, I will comment on the debate between epiphenomenal dualists and their critics, including subject causationists, with a view to identifying: (1) where the debate turns on the assumption of causal closure; and, (2) where the defense of epiphenomenalism would fail for other reasons.

* * *

The situation as seen by the epiphenomenal dualist can be depicted by a diagram like the following adapted from Robinson (2004, 161).

$$\begin{array}{c} Q_1 \\ \uparrow \\ SI \rightarrow P_1 \rightarrow P_2 \rightarrow P_3 \rightarrow P_4 \rightarrow P_5 \rightarrow P_6 \rightarrow P_7 \ldots P_n \rightarrow B \end{array}$$

(Diagram 1, Qualia Epiphenomenalism)

In this diagram, SI represents a Sensory Impact on the nervous system; for example, photons striking the retina. $P_1$ through $P_n$ represent something physical, physical phenomena or neural events, involved in a sequence of causal events (represented by the arrows). B represent the behavior caused by SI via that sequence of causal events. The sequence, $SI \rightarrow P_1 \rightarrow \ldots P_n \rightarrow B$, is completely deterministic.

$Q_1$ represents an experiential phenomenon or qualitative event such as might follow SI. We might take $Q_1$ to be the qualitative event in which Mary first experiences phenomenal redness after being released from captivity. If so, we may then take B to be the report that Mary makes while the experience is still ongoing:

**[MR-1] I am now experiencing *tomato red!***

Apparently, Robinson is only concerned with defending the claim that a causal explanation for B would not involve $Q_1$. There is no downward pointing arrow from $Q_1$ (or any other experiential event) to any physical event.

As a subject causationist, I maintain there is also something else to discuss, the role of the experiencing subject. This is not depicted in any of Robinson's diagrams that I have noticed.

In the following diagrams depicting the subject epiphenomenalist's position and the subject causationist's position, $Q_2$ represents the experiential phenomena or qualitative events associated with decision-making. In Mary's case, $Q_2$ represents her deliberations over whether to report anything at all; and, if she chooses to issue a report, what to say and whether to email or tweet her report to her friends.

$$\begin{array}{ccccccccccccc} & & & & Q_1 & & & & Q_2 & & & & \\ & & & & \uparrow & & & & \uparrow & & & & \\ SI & \rightarrow & P_1 & \rightarrow & P_2 & \rightarrow & P_3 & \rightarrow & P_4 & \rightarrow & P_5 & \rightarrow & P_6 \rightarrow P_7 \ldots P_n \rightarrow B \end{array}$$

(Diagram 2, Subject Epiphenomenalism)

$$\begin{array}{ccccccccccccc} & & & & Q_1 & & & & Q_2 & & & & \\ & & & & \uparrow & & & & \uparrow\downarrow & & & & \\ SI & \rightarrow & P_1 & \rightarrow & P_2 & \rightarrow & P_3 & \rightarrow & P_4 & \rightarrow & P_5 & \rightarrow & P_6 \rightarrow P_7 \ldots P_n \rightarrow B \end{array}$$

(Diagram 3, Subject Causationism)

We've all had experiences of apparently deliberating between two or more alternatives. We've all experienced apparently making a decision and apparently initiating action in consequence of our decisions. But, to the subject epiphenomenalist, the presence of $Q_2$ makes no more difference than the presence of $Q_1$ – note the presence of an arrow pointing upward from $P_6$ to $Q_2$, and the absence of any downward pointing arrow. The sequence, SI → $P_1$ → … $P_n$ → B, is still completely deterministic.

If the subject who experiences deliberating and decision-making – Mary at $Q_2$ – makes a causal contribution to the issuance of her report, it would falsify the claim that only physical causes contributed to the physical effect, B. The qualia

epiphenomenalist may still be able to show that qualia are epiphenomenal; but, if the subject is causally effective, the resulting position seems much more like interactive dualism than epiphenomenal dualism.

### §4.5.4.3.1 The Absence of the Subject

> Suppose someone asks, “Since you are an epiphenomenalist, how do you know that you're not a Zombie?” The answer is, if I am a Zombie, I'm a Zombie; I have no phenomenal consciousness, my overt words about my experiences (necessarily overt, because if I'm a Zombie I haven't got any others) are false, and, of course, I don't know I'm not a Zombie. But if I am not a Zombie, then ***I do have experiences***, and my silent soliloquy about them is true, and caused in a way that makes its truth nonaccidental, and ***I know I have experiences***. And, by the way, I do have experiences. (Robinson, 2004, 172-173 (emphasis supplied))

After reading the “Silent Soliloquy” section of Robinson (2004, 171-173), I emailed Robinson to ask about the ontological status and role of the experiencing subject. His reply was somewhat cryptic; but, indicated that the experiencing subject exists, is not identical to its brain and is not epiphenomenal.[57]

This is the conjunction of claims that constitutes subject causationism; so, the absence of $Q_2$ (and the downward pointing arrow) from his diagrams is puzzling.

Once Mary is released from achromatic confinement and experiences *tomato red*, she reports “I am experiencing *tomato red!*”. If Mary is a qualia epiphenomenalist she is committed to denying that the experiential phenomenon, *tomato red*, caused her report. But, if Mary is a subject causationist rather than a subject epiphenomenalist, she doesn't have to explain how a brain event caused her report about experiencing *tomato red* in a way that made the report true. She has a simple retort, “I caused my report”.

Consequently, I ask the reader for this indulgence. Let us assume that I am evaluating the philosophy of someone who is both a qualia epiphenomenalist and a subject epiphenomenalist as I define those terms but whose position is otherwise as may be inferred from Robinson's cited works. In particular, I will assume that the subject epiphenomenalist draws all the same diagrams that Robinson draws, either leaving out $Q_2$ entirely or having only an upward arrow pointing to it.

### §4.5.4.3.2 Being Determined to Reject Determinism

Some objections to epiphenomenal dualism are based on the allegation that some claims that epiphenomenal dualists want to make are self-defeating, self-undermining or *self-stultifying*, the technical term for this (alleged) predicament.

I will consider more traditional self-stultification objections to epiphenomenal dualism below. At the moment, my argument is not that epiphenomenal dualists undermine their own claims; but, that they undermine their ability to contest my

---

57 However, he took pains to indicate that, on most accounts of what the experiencing subject is, he would deny the existence of an experiencing subject of that kind. I don't blame him for the caveat. With so many accounts of what the experiencing subject is, on most accounts of what the experiencing subject is, I would deny being one of those.

rejection of their epiphenomenal determinism.

Consider the following claim

> **[DRD-1] I am determined to reject determinism.**

Given an admittedly unusual interpretive convention, [DRD] is guaranteed to be true whenever I assert it.

*What is that interpretive convention?*

In English, “determined” has two distinct but equally relevant meanings. It could mean “resolute or “staunchly persevering” ($\text{determined}_r$). Alternately, it may mean something like “predetermined, predestined or controlled by causal influences” ($\text{determined}_c$).

Someone who is $\text{determined}_c$ lives life as if it were a movie to which the script has already been written.

Someone whom I consider to be the archetypal example of $\text{determination}_r$ was Douglas Mawson, an Australian Antarctic explorer. In 1912, Mawson was a member of a 3 man team dog sledding their way across a glacier when disaster struck. One man, six of the dogs and much of their provisions were lost. The two remaining men, Mawson and Mertz, were 300 miles from base camp. There would be no rescue unless they made it back to base camp before their ship departed to avoid the Antarctic winter.

Short of food, Mawson and Mertz sacrificed their remaining dogs to feed themselves and the other dogs. Metz died en route, probably from hypervitaminosis A due to eating too much dog liver.

> Mawson continued the final 100 miles alone. During his return trip to the Main Base he fell through the lid of a crevasse, and was saved only by his sledge wedging itself into the ice above him. He managed to climb out using the harness attaching him to the sled.
>
> When Mawson finally made it back to Cape Denison, the ship *Aurora* had left only a few hours before. It was recalled by wireless communication, only to have bad weather thwart the rescue effort. Mawson and six men who had remained behind to look for him wintered a second year until December 1913. (Wikipedia, 2016-09-21)

In my view, Mawson was resolute, persevering in the face of extreme adversity, $\text{determined}_r$ rather than $\text{determined}_c$ to survive.

Now, in stating [DRD-1] I am using “determined” as if its meaning was in a quantum-like superposition of both relevant meanings, $\text{determined}_r$ and $\text{determined}_c$. In a sense, I won't truly know what [DRD-1] means until all the facts are known, at which point the superposition of alternate meanings will “collapse” to whichever meaning fits the facts.

At the moment, of course, we don't know all the facts; consequently, we can only reflect on the rhetorical situation in which we find ourselves.

In my view, I made a choice to reject determinism as a philosophy; and, I've persevered in that choice, resolutely opposing deterministic theories concerning the brain/subject relation – exactly as I'm doing now. I claim that I am

determined$_r$ rather than determined$_c$ to reject epiphenomenal determinism.

According to the subject epiphenomenalist, my brain caused all the decisions it made related to writing this report of *its* decision – not my decision – to reject determinism. My brain caused my fingers to pound on the keyboard to complete this document. It chose all the thoughts that have been written and are being written, all the claims that have been and will be made. It made all the editorial decisions that sent some passages to the great bit bucket in the sky. It even chose the font in which this document is being presented to you, the reader.

My brain caused it all. I had nothing to do with it except as a silent witness who, for reasons that remain obscure, my brain deceives. According to the subject epiphenomenalist, I (the experiencing subject) am made to think that I made each of the innumerable decisions relating to completing this document. I experienced them all as being my decisions; but, none of them were. I experienced as being mine all the deliberations that went into making these decisions; but, rather than actually being engaged in an inner dialogue in which I tried out various thoughts before selecting one of them, I was just reciting lines written by my brain.

I will assume that the subject epiphenomenalist will concede that I experience feelings of being implacably dedicated to the cause of opposing determinism – Mawson is my muse, remember! However, to the subject epiphenomenalist all such feelings are just more epiphenomenal qualia which have no effect on the outcome, my completion of this document. The subject epiphenomenalist holds that my brain deceived me into thinking that its philosophy is my philosophy.

In short, the subject epiphenomenalists asserts “I am determined$_c$ to reject determinism” whereas the subject causationist says “I am determined$_r$ to reject determinism”.

Suppose that the subject epiphenomenalist tries to explain what he or she thinks is my brain's mistake by saying that my brain made an error while processing information; perhaps, it had faulty or incomplete information and/or operated with faulty logic.

I simply reply “My brain is making me reject your claim because I must reject your claim”. If determinism is false, I am rejecting determinism because I must reject false philosophical doctrines and all arguments offered on their behalf. On the other hand, if determinism is true, my brain has determined that I will report “I reject your arguments for epiphenomenalism because they are question begging or for other reasons” and I must do what I am made to do. Either way, I reject determinism because I must reject determinism.

If the subject epiphenomenalist objects that I've not given my reasons for believing that I have freedom of choice, I can simply reply “I've chosen to believe that I have freedom of choice”. If I do in fact have freedom of choice, I have in fact chosen to exercise that freedom rejection determinism and by telling you that I chose to reject determinism. If in fact I do not have any freedom of choice, my brain is making me claim to have chosen to affirm freedom of choice. In that case, I would be wrong; but, it is not my fault that my brain is unwilling or unable to make me believe that I am an epiphenomenal I.

If subject epiphenomenalists object that I have not provided an argument in the least bit convincing to them, subject causationists could make a reciprocal objection to re-establish an impasse. Each side will be able to defend their own perspective to their own satisfaction using arguments unconvincing to the other side.

This state of affairs is a natural consequence of Robinson's response to one version of the self-stultification objection to qualia epiphenomenalism.

> But how should we reply to those who say, in effect, that epiphenomenalism is unsatisfying because to hold it you must think that you have phenomenal consciousness, but that is something you cannot prove. (Robinson, 2004, 169)

Robinson rejects the suggestion that he is required to prove to others that he has phenomenal consciousness (subjective experience or qualia).

> But the fact that I can't prove to *you* that I have phenomenal consciousness is no argument against epiphenomenalism. ... objectivity may sometimes obtain not in proof to another, but in the common ability for each one to prove something to himself or herself. (Robinson, 2004, 170)

In clarifying what this standard of proof means, Robinson acknowledges what the critic is saying; specifically, that epiphenomenal dualists can't prove to their critics that the epiphenomenal dualist has phenomenal consciousness. However, the epiphenomenal dualist has a reply to such a criticism.

> If I could prove to you that I can prove to myself that I have episodes of phenomenal consciousness, I would have proved the latter to you, which we have just seen cannot be done. So it cannot fairly be demanded that epiphenomenalist prove to others that they can prove to themselves that they have episodes of phenomenal consciousness. (Robinson, 2004, 171)

Now, as a recognition of the limitations of rational argument, this stance toward provability seems reasonable enough. I'm willing to assume that, for any experiencing I, I can't prove to the satisfaction of someone else that I experience anything at all. Nevertheless, when confronted with someone who holds that qualia are epiphenomenal and who claims to know that they experience qualia, the critic may reasonably ask the epiphenomenalist to explain how the claim to knowledge is consistent with the claim of epiphenomenalism.

### §4.5.4.4 Epistemic Arguments

The epistemologist, a philosopher on an epistemological journey toward greater understanding, asks two questions: “What do I know?” and “How do I know it?”.

I'll call the first question the *inventory question* because a complete answer would look like an inventory. Among the many answers to the inventory question would be general statements like “I know the experiential phenomena with which I am acquainted” and “I know the qualia I experience” as well as more .

One approach to answering the second question (which I call the *forensic question*) assumes a causal link between experiencing and knowing, the so-called causal theory of knowledge. Thus, a report that I am experiencing pain (or any other experiential phenomenon) “can express knowledge only if the occurrence it

reports causally contributes to the making of the report". (Robinson, 2006, 88)

This creates a problem for epiphenomenal dualists. An experiential phenomenon – a quale in the terminology of qualia epiphenomenalists – is not identical to any physical phenomenon; but, the report about the phenomenon experienced is a physical effect. It seems intuitively obvious that an experiential phenomenon is causing or contributing to causing a physical effect.

> The most powerful reason for rejecting epiphenomenalism is the view that it is incompatible with knowledge of our own minds – and thus, incompatible with knowing that epiphenomenalism is true. (Robinson, 2015)

Epistemic arguments aim to show that epiphenomenalism is self-stultifying; meaning, either (1) that epiphenomenalists claim to know what epiphenomenalism implies they have no way to know; or, (2) that epiphenomenalists admit to believing what epiphenomenalism implies they have no grounds for believing. Hence the alternate term for epistemic arguments, the Self-Stultification Objection, SSO, of which there are any number of variations.

> If these destructive claims can be substantiated, then epiphenomenalists are, at the very least, caught in a practical contradiction, in which they must claim to know, or at least believe, a view which implies that they can have no reason to believe it. (Robinson, 2015)

The incompatibility follows from assuming that somewhere in between sensory impact and behavioral response is a causal process of a type that epiphenomenalism forbids.

> The argument that is given to support the destructive claims is that (i) knowledge of one's mental events requires that these events cause one's knowledge, but (ii) epiphenomenalism denies physical effects of mental events. (Robinson, 2015)

It is safe to assume that sensory impacts and behavioral responses (SI and B in diagram 1) are physical events. Philosophers of various persuasions would agree that $Q_1$ – an occurrence of pain, phenomenal redness or some other experiential phenomenon – occurs somewhere in between the occurrence of SI and the occurrence of B. For anyone who asserts that $Q_1$ is not identical to any of the physical events $P_1 \ldots P_x$ in between SI and B, an occurrence of experiencing $Q_1$ is an experiential event not a physical event. But, if epiphenomenalism is true, $Q_1$ can have no causal influence on the SI → $P_1 \ldots P_n$ → B sequence of physical events.

> So, either we cannot know our own mental events, or our knowledge of them cannot be what is causing the plainly physical event of our saying something about our mental events. Thus, suppose S is an epiphenomenalist, and that S utters "I am in terrible pain." S is committed to the view that the pain does not cause the utterance. (Robinson, 2015)

It is difficult to believe that experiencing pain does not cause knowledge of being in pain and does not cause or contribute to causing any subsequent behavior such as moving away from the source of the pain, taking pain relieving medication or telling someone else about the pain. But, if the pain itself does not cause the report, whether directly or indirectly, how does the epiphenomenalist explain the occurrence of the report about the pain?

Robinson's answer is that a report about $Q_1$ can express knowledge of $Q_1$ if it is caused by the cause of $Q_1$.

> The neural events that cause qualitative events also cause other effects in the brain and it is these further effects of the causes of qualitative events that can serve as inputs to many brain systems, including our language systems, and issue in useful behaviors, including reports of our qualitative events. (Robinson, 2006, 88)

Robinson appears to interpret his reply as a rejection of the causal theory of knowledge; but, in my view, it seems more like a peculiar interpretation of that principle to me. Something that is alleged to be knowledge is still held to have caused the report; the report is still held to be about $Q_1$ rather than the brain event(s) that caused the report; and, having a common cause provides an epistemic warrant for taking the report to be a reliable indicator of what it purports to be about.

#### §4.5.4.4.1 Rebutting the Common Cause Reply

Other philosophers have been skeptical of Robinson's reply, a version of the Common Underlying Cause Reply to the self-stultification objection. (See Moore (2012b) for a critique of the Common Underlying Cause Reply and Moore (2012a) for a critique specifically about Robinson's position. See Robinson (2013a) for his reply to Moore. See De Brigard (2014) for a criticism focusing on Robinson's claim that qualia epiphenomenalists are at no disadvantage relative to identity theorists when it comes to explaining the behavioral response.)

I find these critiques unpersuasive. Philosophers of various perspectives accept that there are correlations between the occurrence of experiential phenomena and the occurrence of physical phenomena. One way of explaining a correlation between two occurrences is to say that they are both caused by the same event; and, until we have empirical research results ruling out this possibility, I don't see the basis for holding that epiphenomenalists can't theorize along these lines. And, of course, once it is conceded that the two occurrences are correlated, it is difficult to deny that one may be a reliable (although maybe not a perfectly reliable) indicator of the other.

Instead of trying to strengthen such critiques, I will pursue two alternative strategies. First, I will consider the virtues of walking through the loophole that Robinson acknowledges in his argument against the causal theory of knowledge. Second, I will draw attention to Robinson's reliance on an equivocation as to the meaning of “knowledge”.

#### §4.5.4.4.2 Through the Loophole

Even if the path from sensory impact to behavioral effect is completely deterministic, it wouldn't follow from that fact alone that the report about $Q_1$ is knowledgeable – expresses knowledge about $Q_1$. To reach that conclusion, it appears that Robinson implicitly assumes that some neural phenomenon, $P_x$, in the diagrammed sequence between SI and B, just *is* knowledge of $Q_1$, the quale being experienced.

However, one can simply deny this assumption, as Robinson goes on to

acknowledge.

> Very strictly speaking, there is one alternative which avoids our rejection of [the causal theory of knowledge] which we have not yet considered. This is the claim that knowledge is a purely mental state which is not connected with behavior in the way that our discussion tacitly assumes. On this gambit all the above remarks about [$P_x$] are simply irrelevant since neither it nor any of its effects are knowledge. This gambit, however, is really a verbal maneuver. (Robinson, 1982, 536)

The claim that supposedly avoids rejecting the causal theory of knowledge is the claim that knowing an experiential phenomenon is a purely mental state … of the experiencing subject, presumably. Since this “gambit” introduces the experiencing subject into the conversation, we might well call it the *subject gambit;* but, whatever it is called, not only is the gambit is playable, it is compulsory. Qualia epiphenomenalists are in *zugzwang* when it comes to the subject of experience. They must take a stand as to the ontological status of the experiencing subject and its involvement in reports about experiencing.

At this point, let us review the possible rhetorical strategies.

The epiphenomenalist holds that experiential phenomena have no causal impact on anything physical; so, just as Robinson indicated, the epiphenomenalist, S, who reports “I am in terrible pain” is committed to the claim that the pain does not cause the utterance of the report.

Very much under-appreciated, in my view, is that S is also committed to the existence of the experiencing subject. A report such as “I am in terrible pain; and, just to be crystal clear, the word 'I' as I used it in reporting 'I am in terrible pain' is a non-referring term because I don't exist in any sense whatsoever” is simply incoherent.

If S is an qualia epiphenomenalist who holds that the experiencing subject exists but is not identical to its brain (or some other physical object, the whole body perhaps), he or she may also hold that the experiencing subject has no causal impact on anything physical (subject epiphenomenalism). S would be committed to the view “The pain I experience does not cause my utterance of 'I am experiencing pain' and I (this experiencing subject) don't cause my utterance either”.

If S is an qualia epiphenomenalist who holds that the experiencing subject exists and is identical to its brain (or some other physical object, the whole body perhaps), he or she may say “The pain I experience does not cause my utterance of 'I am experiencing pain' but I (this experiencing brain) do cause that utterance”.

Now, I hold that the experiencing subject exists, is not identical to its brain but *is* causally effective at least some of the time. As a subject causationist, I account for the occurrence of my report about $Q_1$ by saying “At the time of $Q_2$, I chose to issue the report (after a short period during which I deliberated over whether to report on a pain, an afterimage, something else or nothing at all)”.

### §4.5.4.4.3 Knowing by Acquaintance

Given that I chose to mention the pain I am experiencing, the pain is a necessary condition of my knowing that pain; but, it is not clear that being a necessary condition of knowing *causes* knowing – at least where knowing by acquaintance is concerned. I might say that knowing experiential phenomena by acquaintance depends on the experiencing subject noticing its experience; but, I wouldn't say that noticing my experience *causes* knowing my experience.

Of course, I need the pain to come to my attention; but, this doesn't necessarily mean that the afterimage itself has causal powers sufficient to bring itself to my attention. An advocate of subject causation may concede without self-contradiction that the brain does all the work. I may theorize that I experience when information that is physically instantiated in my brain activity becomes phenomenally instantiated in my experience.

Naturally, the qualia epiphenomenalist is free to reject the possibility that information has two modes of instantiation; but, my point is not that the qualia epiphenomenalist is compelled to agree with the subject causationist. Rather, my point is that subject causationists may agree with qualia epiphenomenalists that the brain, not the pain, is responsible for bringing the pain to the attention of the experiencing I – *this* whatever it is that reports “I am experiencing pain”.

Now, the subject causationist who also advocates qualia epiphenomenalism still needs to evade the causal theory of knowing experiential phenomena; but, this is easily done. In the view of many philosophers of various persuasions, I become acquainted with experiential phenomena as I experience them (as my brain brings them to my attention); but, as discussed earlier, knowing by acquaintance is not universally considered a causal process; and, I am inclined to agree with Chalmers that knowing by acquaintance is *not* a causal process.

### §4.5.4.4.4 Conundrum

Robinson characterizes his approach as a rejection of the causal theory of knowledge; and, he goes to great lengths to reply to objections based on the causal theory of knowledge; but, he turns down the opportunity to sidestep the causal theory of knowledge altogether by taking the experiencing subject's knowledge of experience as knowledge by acquaintance.

It's a conundrum that is only partially explained by Robinson denial that we have knowledge by acquaintance, at least as advocates of knowledge by acquaintance understand that term.

In the course of explaining the relation between experiencing and our judgments about experiencing, Robinson is careful to note that

> … explaining my judgment that I have an F experience by something extraordinarily simple – the mere occurrence of an F experience – requires acceptance of mysterious connections; and no such mysterious connections are needed. (Robinson, 2004, 176)

Robinson acknowledges that the justification of our judgments about our experiences does not involve the usual sense of “justification” which refers “to the logical relation between premises and conclusions” (2004, 176). Instead he

speaks of being justified because

> … in a quasi-moral sense, we are not doing something that is “unjustified” if we judge that we have an F experience – i.e., we are not eligible for criticism as being careless reasoners or uncritical believers if we so judge. (Robinson, 2004, 177)

This quasi-moral sense of being justified in commenting on our experiences may be reminiscent of knowledge by acquaintance; and, Robinson has no objection to that term “*provided* that it is regimented to have no other sense than 'knowledge of phenomenal qualities'” (2004, 177).

However, while Robinson acknowledges that such knowledge is knowledge despite being non-inferential and direct, he believes that the phrase “knowledge by acquaintance” has the potential to be misleading for several reasons, including that “the phrase may suggest that *having qualitative events* is *in itself* a kind of knowledge”.

Now, someone who advocates that we have knowledge by acquaintance of experiential phenomena is saying precisely what Robinson doesn't want to say: that acquaintance is a special epistemic relation – a mysterious connection – between the experiencing subject and the object of its awareness; and, that just experiencing a phenomenal quality *is* knowing that quality. We want to say that when Mary emerges from achromatic confinement and learns what experiencing *tomato red* is like, she knows what experiencing tomato red is like because she *knows* – is acquainted with – *tomato red*, the experiential phenomenon or phenomenal quality.

To focus attention on the issue at hand, let us suppose that Mary, when she first encounters a tomato after her release, reports “I am now acquainted with *tomato red*”. Is her statement true or false; and, how would epiphenomenalists explain their answer?

If Mary does not have acquaintance knowledge, her statement is false; so, whatever brain state is causing her to assert that false claim *lacks* relevant knowledge, contrary to Robinson's claim that the report is caused in a way that makes its truth non-accidental. In cases where epiphenomenalists disagree with the content of the report, is the report caused in a way that makes it false?

If Mary has acquaintance knowledge, her statement is true; so, the next questions would be (1), whether that acquaintance knowledge causally contributes to the issuance of her report; and, (2), whether that acquaintance knowledge is expressed in the content of the report.

Robinson denies the possibility of knowing by acquaintance; so, he must rely on some other way of knowing to justify claiming that we have direct, non-inferential knowledge of phenomenal qualities; but, it's not clear just what that way of knowing is. Further consequences of denying the possibility of knowing by acquaintance are that Robinson must deny that acquaintance knowledge causally contributes to the issuance of reports about experience and that acquaintance knowledge is expressed in the content of such reports.

These point will become clearer as we compare Robinson's argument for being able to prove to himself that he is not a zombie with Chalmers' argument for the

same claim.

Chalmers concedes that other people might not be able to distinguish him from his zombie twin; because, from the third person point of view, he and his zombie twin are indistinguishable. However, the first-person point of view provides a different epistemic perspective.

> Finally, there is a persistent refrain that comes up in these situations: "But your zombie twin would say the same thing!" If I say I know I am conscious, it is noted that my zombie twin says the same. If I say my belief is justified by my immediate acquaintance with experience, it is noted that my zombie twin says the same. To this, the answer is again, "So what?" At most this shows that from the third-person point of view, my zombie twin and I are identical, so that you cannot be certain that I am conscious; but we knew this all along. But it does nothing to imply that from the first-person view, I cannot know I am conscious. From the first-person point of view, my zombie twin and I are very different: ***I have experiences***, and he does not. Because of this, I have evidence for my belief where he does not. Despite the fact that he says the same things I do, I know that I am not him (though you might not be sure) because of my direct first-person acquaintance with my experiences. This may sound somewhat paradoxical at first, but really it is simply saying the obvious: our experience of consciousness enables us to know that we are conscious. (Chalmers, 1996, 198 (emphasis supplied))

Robinson say this:

> Suppose someone asks, “Since you are an epiphenomenalist, how do you know that you're not a Zombie?” The answer is, if I am a Zombie, I'm a Zombie; I have no phenomenal consciousness, my overt words about my experiences (necessarily overt, because if I'm a Zombie I haven't got any others) are false, and, of course, I don't know I'm not a Zombie. But if I am not a Zombie, then ***I do have experiences***, and my silent soliloquy about them is true, and caused in a way that makes its truth nonaccidental, and ***I know I have experiences***. And, by the way, I do have experiences. (Robinson, 2004, 172-173 (emphasis supplied))

Although it looks like they are saying virtually the same thing, “I have experiences” vs “I know I have experiences”, they mean something very different. For Chalmers, it is knowledge by acquaintance that underwrites claims such as “I have experiences” and “I know I am conscious”. Robinson denies knowledge by acquaintance; so, the claim that we know we have experiences must rest on some other way of knowing.

This way of knowing is not named; but, we know that the report (in this case a silent soliloquy) is caused in a way that makes the knowledge claim “I have experiences” non-accidentally true. So, let us proceed to connect the dots.

For Robinson, the brain activity (or some of it) that generates experiential phenomena also counts as knowledge of experiential phenomena – phenomenal qualities in his terminology.

To address this question, I want to revisit the passage in which Robinson introduces the subject gambit.

> Very strictly speaking, there is one alternative which avoids our rejection of [the causal theory of knowledge] which we have not yet considered. This is the claim that knowledge is a purely mental state which is not connected with behavior in the way that our discussion tacitly assumes. On this gambit all the above remarks about [$P_x$] are

> simply irrelevant since neither it nor any of its effects are knowledge. This gambit, however, is really a verbal maneuver. (Robinson, 1982, 536)

I do not take as harsh a view toward the theory that a brain state can constitute knowledge (of some sort) as Robinson presumes an advocate of knowledge by acquaintance would take. I accept the possibility that brain activity that instantiates information may constitute knowledge of some sort; but, I reject two other notions that seem implicit within Robinson's position:

1. If nervous system activity instantiating information counts as knowledge, it counts as the experiencing subject's knowledge rather than the brain's knowledge; and, more specifically,
2. Nervous system activity instantiating information about the retinal response to incoming photons counts as the experiencing subject's knowledge of the experiential phenomena to be caused by whatever brain activity occurs when the signal from the retina arrives in the brain.

This distinction between the knowledge I have and the knowledge my brain has is rooted in the claim of brain/subject non-identity.

Among those who assume that brain activity which instantiates information counts as some kind of knowledge, the brain/subject identity theorist may assume or conclude that the experiencing subject has whatever knowledge the brain has because the experiencing I and its brain are one and the same, self-identical item.

The issue is more complicated for the brain/subject non-identity theorist. I will explore the issue in the next subsection by considering a simple example, eye moving events.

#### §4.5.4.4.4.1 Do I Move My Eyes?

Let us return to the questions that drives us along the epistemological journey: *What do I know; and, How do I know it?*

Suppose that, taking inventory, I ask myself: *Do I know how to move my eyes?*

My answer is that, strictly speaking, I do not know how to move my eyes.

When I began writing this thesis, I did not even know how many eye muscles I have; but, I was able to move my eyes anyway. I now know that each eye has six muscles that control its movements; but, my ability to move my eyes does not depend on me having that knowledge or any other knowledge that I might acquire by reviewing the scientific literature. Even now, I don't know which nerves are responsible for stimulating each eye controlling muscle; so, my ability to move my eyes (assuming I have such an ability) doesn't depend on having knowledge about neurophysiology.

*Am I able to move my eyes?*

Strictly speaking I do not move my eyes. I am not claiming the power of telekinesis; hence, I deny reaching out (from wherever I am) to directly trigger the eye muscles into action.

*Should I conclude that I am unable to move my own eyes merely because I do not move my eyes by having a direct, telekinetic impact on my eye muscles?*

I think not. As I experience eye-moving events, it seems clear that, absent special circumstances, I generally do not intend an eye movement as such. However, if I intend some action that requires an eye movement, my eyes usually move in accordance with my intent.

I assume, as a forensic inference to the best explanation for how that happens, that my brain knows how to move my eyes; and, that my brain moved my eyes in accordance with my intent. As a further reflection on these inferences, I might conclude that my brain is intent responsive; at least, to some extent.

In these circumstances, it is entirely reasonable to attribute one sort of knowledge to the brain (knowledge of how to move my eyes) and knowledge of a different sort to the experiencing I (knowledge of how to make my brain move my eyes for me).

Nevertheless, I'm also willing to say, in a looser but still appropriate sense, that I moved my eyes because I triggered the brain activity that resulting in the eye movement. Similarly, using a suitably relaxed sense of knowing how, I could claim to know how to move my eyes because I assume that my brain knows how to move my eyes and I know that intending actions that require an eye movement will make my brain move my eyes for me.

This is not a special case involving innate knowledge built into the nervous system. The same situation occurs with learned movements. Very early in my career as a Star Trek fan, I learned how to give the Vulcan salute. I have no idea which muscles are moved by which nerves to accomplish this gesture. I only know that, when I intend to perform the Vulcan salute, it occurs as intended. I assume my brain knows how to implement my intention and responds accordingly.

The difference between the subject epiphenomenalist and the subject causationist does not consist in whether we attribute knowledge to a brain state of which the subject is unaware. We're both willing to do that, at least in some circumstances.

The difference is that the subject causationist claims “I triggered brain activity by intending something my brain knows how to do”. The subject epiphenomenalist denies this, instead claiming that some brain activity caused both the illusion of intending action and the action corresponding to the illusionary intent. Furthermore, this brain activity – of which the experiencing subject is unaware, mind you – is supposed to be the subject's knowledge of the experiential phenomenon (of which he or she *is* aware).[58]

---

58 In earlier commentary on the debate over Jackson's KA and my (hopefully) improved version, KA:TNG, there was a strong concern about the validity of arguments based on premises such as “I am aware of experiential phenomena” and “I am not aware of physical phenomena in the brain”. A conclusion such as “Experiential phenomena are not identical to physical phenomena in the brain” is (according to identity theorists) vulnerable to the objection that experiential phenomena could be identical to physical phenomena without me knowing that to be the case. Objections of that sort are not applicable here because the epiphenomenal dualist already

This is the point at which I contest Robinson's attempt to treat brain activity as knowledge. I'm willing to assume arguendo that brain activity which instantiates information is a kind of knowledge; but, if I am not identical to my brain, it does not follow (except in some derivative sense) that I have whatever knowledge my brain has; and, it certainly doesn't follow that my brain has whatever knowledge I have as an experiencing subject.

We might each have knowledge; but, different kinds of knowledge. In particular, I could concede that the my brain has knowledge in the sense of having brain activity that instantiates information while denying that such is knowledge is knowledge of experiential phenomena.

Let's take a situation more conducive to a discussion about qualia epiphenomenalism. Mary is released from her achromatic confinement, looks at her tomato, experiences *tomato red* for the first time, deliberates for a bit and reports “I am now experiencing *tomato red*”.

It is entirely likely that light reflecting off the surface of the tomato carries information about the surface of the tomato – that it absorbs light at wavelengths not present in the reflected light. When the light strikes the retina it is transduced into a signal instantiating information about the responses of the rods and the cones of the eye to the light reflecting off the tomato; and, it seems quite reasonable to assume that this information is conveyed to the brain where, possibly after further processing, it causes or contributes to causing an occurrence of an experiential phenomenon – *tomato red*, in this case.

The subject epiphenomenalist is free to say that some brain state, $P_x$, as we've been calling it, causes both the quale Mary knows as *tomato red* and her report about it. The subject causationist will naturally disagree; but, that disagreement has nothing to do with whether $P_x$, a state of information instantiation, is also considered a state of knowledge.[59]

I may not speak for all subject causationists; but, I'm willing to assume that a brain state that instantiates information is a state of knowledge (of some sort); but, I find it impossible to believe that a signal traveling up the optic nerve constitutes knowledge of the experiential phenomena the brain will eventually generate after receiving that signal. That signal carries information about the retinal response to incoming photons; so, assuming it constitutes knowledge of some sort, it seems more plausible to conclude that the signal leaving the retina counts as knowledge of the retinal response to incoming photons.

If information about the retinal response to incoming photons is all the information the visual cortex receives, it seems reasonable to conclude that it does not have knowledge of the experiential phenomena that it somehow generates in response to receiving information about the retinal response to some stimulus.

---

concedes the non-identity of experiential and physical phenomena.

59 I'm not willing to say that *any* information storage counts as knowing. My hard drive doesn't know anything about philosophy even though it is storing this entire document as well as numerous PDF files of articles that I've referred to while writing it; but, I may be persuaded to cut my brain more slack than I'm willing to give my hard drive.

### §4.5.4.4.4.2 My Brain Has Information; Do I Have Knowledge?

Assuming that Mary acquires new knowledge as a result of her encounter with *tomato red*, it is natural to wonder what sort of knowledge she acquires. Is Mary's knowledge knowledge by an experiencing subject able to know by acquaintance; or, is it knowledge by a brain able to instantiate information? Does she have both kinds of knowledge?

I'm willing to assume that neural activity instantiating information constitutes knowledge of some sort. What I'm not willing to assume is that the knowledge argument is about acquiring that sort of knowledge. Mary, according to the terms of the KA, already has complete knowledge of electrochemical activity in the brain; so, she already knows about the signals resulting from retinal activity. In my view, knowledge arguments are about whether the experiencing subject acquires knowledge by acquaintance.

The qualia epiphenomenalist who denies knowledge by acquaintance may argue: Mary acquires knowledge of *tomato red*; there is no knowledge by acquaintance; so, her knowledge must be acquired in some other way. At this point, an abductive leap would be required to hypothesize that the knowledge Mary acquires upon her release is present in the form of brain activity. But a qualia epiphenomenalist taking this position would still need an argument showing that we are not acquainted with experiential phenomena or that we gain no knowledge from acquaintance with experiential phenomena.

Robinson argues that we experience qualia without being acquainted with them; but, it is not clear how the experiencing subject can have knowledge of the qualia it experiences if it is not acquainted with those qualia. The knowledge that makes a report about experiencing qualia knowledgeable is, according to Robinson, the knowledge constituted by brain activity; but, it is not clear why brain activity consisting of neural phenomena instantiating information about retinal activity should count as the experiencing subject's knowledge of experiential phenomena.

The striking thing about knowing experiential phenomena by acquaintance is that the subject is aware of the phenomenon in question. But the experiencing subject need not and typically would not be aware of the neural phenomena that instantiated information about the retinal response to incoming photons.

To have a complete theory of experiencing, qualia epiphenomenalists who deny that we have knowledge by acquaintance with experiential phenomena need to explain how physically instantiating information about retinal activity counts as knowledge *I* have rather than as knowledge (in some other sense) that my brain has. I can become acquainted (or re-acquainted) with *tomato red* without knowing anything about retinal activity or anything about any brain activity associated with experiencing *tomato red*.

In any case, whether $P_x$ is considered knowledge or information storage, the *qualia* epiphenomenalist can't claim that $P_x$ causes the report about $Q_1$ (in this case Mary's instantiation of *tomato red*) without first assuming or concluding that *neither* $Q_1$ *nor* $Q_2$ (in this case Mary's deliberations over what, if anything, to report) had anything to do with it.

### §4.5.4.5 Twitching on the Verge of Sleep

Let us assume that at least some qualia epiphenomenalists are willing to affirm subject epiphenomenalism. Let's assume further that qualia epiphenomenalists are able to convince themselves that they have adequately defended themselves from all rational arguments against their position. There is still the possibility that there is empirical evidence bearing on this question.

As long as I have the causal power that I, consciousness qua subject, appear to have, the power to decide what (if anything) to say about my afterimage, I'd prefer that the afterimage itself be epiphenomenal. Just thinking that experiential phenomena might be doing things in/to/with my brain without my knowledge or consent leaves me with a creepy feeling.

And therein lies a novel argument against qualia epiphenomenalism: creepy things happen.

Occasionally, while I am on the verge of sleep, my body will “twitch”. In a typical experience, I'll find myself in a dream standing motionless. Someone throws something, most often a ball, toward me. In the dream, I reach out to catch the ball. My physical arms move just as the dream depicts my dream arms moving; and, that wakes me up.

I assume that the twitch is an instance of what is technically a myoclonic jerk but which is more popularly known as a hypnic or hynagogic jerk or a sleep start. However, I don't have the falling sensation usually associated with a hypnic jerk.

I take the experiential component of a twitch event to be a dream; but, it may actually be hypnogogic imagery. What's important at the moment is that it seems to me experiencing such an event that my body moves – all by itself – in response to the visual imagery. I did not intend to move my arms; but, they moved anyway! Perhaps my brain is not just intent responsive. Perhaps it is also qualia responsive.

*How is it possible for my body to twitch in response to epiphenomenal imagery?*

Perhaps the twitch occurs because my brain begins generating imagery before it fully paralyzes my body in preparation for dreaming; but, even if such an explanation accurately describes the psychophysical event(s) I experience, it does not begin to explain why evolution equipped the brain with a means to prevent the sleeping body from moving in response to epiphenomenal imagery.

As creepy as it is to twitch on the verge of sleep, I consider myself fortunate not to suffer from a more extreme failure of the process by which the brain paralyzes the body when it cycles into the REM sleep state.

> REM behavior disorder (RBD) is a rare parasomnia and very often is misdiagnosed. It is characterized by the intermittent loss or impairment of REM sleep atonia and by the appearance of elaborate motor activity associated with vivid dream-enacting behaviors. (Boufidis, 2008, 1)

Taking REM Behavior Disorder as dream-enacting behavior makes epiphenomenalism too implausible to believe. I can think of no reason why evolution went to the trouble of equipping the brain with the means to paralyze

the body so that it wouldn't react to epiphenomenal qualia.

The qualia epiphenomenalist may reply that the brain generates the experiential phenomena of the dream as well as the signals that are sent to the limbs, signals to which the body would respond but for the paralyzing signal also sent. I concede that such an explanation may seem convincing to epiphenomenalists; but, if I am an epiphenomenon, I am an epiphenomenon who is unconvinced by such explanations and who denies being an epiphenomenon. To me it seems more likely that the brain has the function of paralyzing the body to render REM state experience epiphenomenal when, but for REM state atonia, it would have causal effects.

It seems that we have once again achieved an impasse. Any attempt to explain how I could be an epiphenomenon explains why I reject epiphenomenalism.

#### §4.5.4.6 Assessment of Epiphenomenalism

Epiphenomenalism is an unpopular theory.

> Many philosophers take it for granted that epiphenomenalism is obviously a dead end for an understanding of the human mind and its relation to the physical world and nothing but a counterintuitive theory of last resort. (Pauen et al.; 2006, 7)

One naturally wonders why anyone would hold it.

> [Epiphenomenalism is] a theory of last resort – one into which people are pushed by the sense that all the alternatives are even less plausible. (Rudd, 2000, 60)

Epiphenomenalists may well agree that their theory is counterintuitive; but, I doubt that many epiphenomenalists think of their own theory as a theory of last resort; although, a non-identity theorist with a prior commitment to the CCP may find that accepting epiphenomenal dualism is the only way to avoid interactive dualism.

*But, why accept the CCP in the first place?*

Epiphenomenalists argue for their position on the assumption of causal closure, an assumption for which no argument is given; and, that may make perfect sense in a debate confined to those who accept causal closure; but, if causal closure is not true, all those arguments are unsound even if not fallacious.

With no good reason for accepting the CCP in the first place, epiphenomenalism is not a real option.

## §4.6 The Discrepancy Thesis

Physicalism, the philosophy, is supposed to remain consistent with physics, the science, interpreted broadly enough to include all physical sciences. However, there is a problem. I call it *the Discrepancy Thesis*.

> **[DT] There is a discrepancy between what physical scientists say about the physical world and what physicalist philosophers say about it.**

Specifically, most physicists adopt a free will postulate while most physicalist

philosophers adopt some version of the causal closure principle.

*What are we to make of this?*

Even if it is true that most physicists adopt a free will postulate for theoretical physics, the popularity of this postulate among physicists does not prove that the postulate is true; only that it is assumed by most physicists – making the free will postulate (or its denial) an element of the philosophy of physicists.[60]

Nevertheless, it should be noted that deterministic physicists do not defend their opposition to the free will postulate by citing Kim's causal closure principle or Searle's causal reduction principle. If the question is to be settled, it will have to be settled empirically. Until then, we must chose between competing philosophical intuitions. One may reject the causal closure principle in favor of a free will postulate; and, unless one's viewpoint is inconsistent with empirical findings, it should still count as physicalism.

I choose to adopt a free will postulate for my philosophical perspective. If I am, in fact, free to choose from among the options available to me at the time of choice; then, I'm right to adopt a free will postulate. If we are not free to choose, it's not my fault that I've made an incorrect choice with respect to freedom of choice. Determinating influences make me reject both determinism and epiphenomenalism. Or, perhaps, they were random influences; but, in the 40+ years since I first began thinking about philosophical issues, I've wondered about freedom of choice many times. And, I've always come to the conclusions that we have and make choices. I reject the possibility that for 40+ years, random influences have consistently required me to reject the possibility that my decisions have been due to random brain events rather than deliberate choices.

Adopting a free will postulate for philosophy has consequences. One must confront the fact that philosophers currently don't have a good way to explain the possibility of intentional action without substance dualism; so, one must embrace substance dualism, adopt some version of mysterianism or find some other explanation.

Second, if philosophers claim that they can explain experiential phenomena using only the tools that scientists use to explain physical$_1$ phenomena, they will

---

60 There are some physicists (e.g. 't Hooft, 2007) who reject the free will postulate in favor of what John S Bell called “superdeterminism” – a conspiratorial determinism in which the universe arranges for physicists to perform experiments that give them a false impression of the laws of the universe. Interesting questions for experimental philosophers and psychologists would include: (1) whether those philosophers as well as physicists who favor the free will postulate over a conspiratorial determinism do so because of a belief analogous to Einstein's belief that nature may be subtle but not malicious; and, (2) how those who favor a conspiratorial determinism explain how they understand the truth. It's not much of a conspiracy if everyone can discover it; but, if only a chosen few can perceive the truth, there are special people in the world. Another dissenter, Dugarte (2014), claims to have proven that the evolution of the universe is univocally determined; and, therefore, that free will does not exist. This, too, creates the special people problem. Some people accept determinism and some people reject it. If determinism is true it would mean that the laws of the universe determine that some people will be correct about determinism and that some people will be incorrect. What physical phenomenon explains this (presumably mental) difference? Is it a genetic difference?

have to do so without relying on the principle of causal closure of the physical. It is not a finding of physics.

Third, we've taken a step toward a radical revisioning of the philosophy of consciousness. In days gone by, it was possible to think that scientists would someday explain consciousness on the basis of physics; but, arguably, physics now presupposes consciousness. If that is the case, a conceptual revolution is in the making.

## §5 In Search of Scientific Explanations

In his classic paper, *Sensations and Brain Processes*, J. J. C. Smart laments,

> So sensations, states of consciousness, do seem to be the one sort of thing left outside the physicalist picture, and for various reasons I just cannot believe that this can be so. That everything should be explicable in terms of physics … except the occurrence of sensations seems to me to be frankly unbelievable. (Smart, 1959, 142)

I share Smart's sentiment to some extent; but, I reject what I take to be his expectation: that everything will eventually have a *reductive* physical explanation. In my view, incredible is hoping that scientists will someday prove that experience is nothing more than something else.

What expectations are appropriate for our endeavor to understand the true nature of the brain/experience relation? Clearly, we must have some criterion of success against which to evaluate claims of success; but, it's not obvious where to set the bar. Arguably, Chalmers set the bar too high. “For a satisfactory theory, we need to know more than which processes give rise to experience; we need an account of why and how.” (1995, 207)

I doubt that scientists will ever answer the “why” question empirically; but, I suspect that few will be greatly distressed by that failure. The question “Why is there experience rather than no experience at all?” may be like the question “Why is there something rather than nothing?”. There may be no answer other than “there just is”.

On the other hand, I believe that we are practically guaranteed an answer to the “which” question, eventually. Scientists have been very successful at identifying the brain activity associated with a given experiential phenomenon; and, one can readily agree with Chalmers that there is “good reason to believe that there is a lawful relationship between physical processes and conscious experiences”. (Chalmers, 1996, 127)

We can easily imagine phenomenologists diligently cataloging various aspects of first-person phenomenology, and heterophenomenologically inclined neuroscientists examining those reports for clues that would help identify the brain activity correlated with salient features of those reports. So, we can reasonably expect a scientific account of the brain to include statements about what happens in the brain *and* statements about what experiential phenomena, if any, typically accompanies a given kind of physical phenomenon. It may be observed, for example, that incoming signals of a certain kind will cause C-fibers to start firing *and* that subjects will typically report experiencing pain when that happens.

At some level of generality, correlated statements become psychophysical laws that say when this happens that will be experienced or when this is experienced that will be happening. It's quite likely, in my view, that scientists will eventually discover objective criteria by which to decide whether some specified brain activity is likely or unlikely to be accompanied by experiential phenomena. This would satisfy one of the criteria that Chalmers gives for a truly final theory. We will have an account of everything that happens. We will know *which* experiential

phenomena and *which* physical phenomena are correlated; and, the psychophysical laws connecting the two will tell us something about the nature of the relationship.

However, a theory correlating physical and experiential phenomena will not, by itself, explain *how* experience arises from or *how* it is generated by whatever it is correlated with.

It is always possible that, somewhere along the way, scientists will discover how experience is generated and from what; but, there is no guarantee that this will occur. There is always a risk that we will someday make a choice: to savor the mystery that remains after scientists have explained whatever they are capable of explaining; to add extraneous ingredients to the scientific account to get the type of explanation we want; or, to adapt our expectations to what a scientific explanation of experience provides.

## §5.1 Opting for Mysterianism

My position could be described as a moderately optimistic mysterianism.

I'm a mysterian in that I decline to assume my way across an epistemological gap; preferring instead to cultivate awareness of the profound mysteries of consciousness *qua* experience and of consciousness *qua* subject of experience.

I'm optimistic in that, unlike McGinn who already knows enough to conclude that humans will never be able to know the true nature of the brain/consciousness relations, I anticipate that scientists will eventually a produce a single well-supported theory of the brain/experience and brain/subject relations by experimentally falsifying other theories. Nevertheless, I acknowledge a serious risk that, some bittersweet day in the future, the philosophers of that day will have good reasons to look back and say that McGinn was right all along. But, today is not that day. It's too soon to say that understanding the brain/experience and the brain/subject relations is inherently, now and forever, beyond our ken.

Nevertheless, I am not as optimistic as Nagel (1998) who describes his position as “fairly close to that of Colin McGinn, but without his pessimism”. Nagel believes that we will not understand the brain/experience relation until a conceptual revolution gives us a way to understand that the relation is a necessary identity. I agree that a conceptual revolution is needed; but, in my view, the good ship *Identity* set sail long ago … and then sank after being shipwrecked by an afterimage.

It is currently popular to assume either that experience is identical to or that experience arises from the brain activity with which it is correlated; but, until physicists empirically falsify the possibility that they must postulate an immaterial consciousness to collapse the quantum wavefunction, neither physicists nor philosophers can rule out the possibility that such an immaterial but causally efficacious consciousness may also play a role in explaining experiential phenomena.

That said, my reasons for whatever degree of optimism I have are that we have considerable flexibility in defining what counts as a scientific explanation and in

adapting our expectations accordingly; and, that we may simply choose to live with a mystery rather than embrace an answer just to have an answer.

## §5.2 Opting to Adapt Our Expectations

It may be that we will have to learn to savor the mystery of consciousness forever; but, before making (or rejecting) that choice, let's ask ourselves just what would be included in the scientific account of the world. It may be that we would not have to lower our expectations too much. Not only would that reduce the extent of the mystery that remains unexplained by science, it would also reduce the motivation to add non-scientific elements to the scientific account.

The type of theory that would be consistent with assuming that experience is fundamental is exactly the type of theory that scientists are already working on, a theory of physical and experiential phenomena that describes correlations between and describes their interactions by reference to psychophysical laws.

The point is that scientists do not have to assume that experience is a fundamental fact about the world before they begin their investigations. They just have to assume that the brain/experience relation is a topic worthy of scientific investigation; *and, actually conduct those investigations*. The assumption that experience is fundamental, a primitive (irreducible or unexplained) element in the ontology of physics, would be justified either as an inference based on the empirical failure of theories that assume that experience is not fundamental or as a choice based on the recognition that theories that achieve an explanation only by adding something extraneous to the scientific account are questionable.

## §5.3 Opting to Add to the Scientific Account

In Dreams of a Final Theory, physicist Steven Weinberg, suggests that even a so-called theory of everything might not explain consciousness.

> It is not unreasonable to hope that when the objective correlatives to consciousness have been explained, somewhere in our explanations we shall be able to recognize something, some physical system for processing information, that corresponds to our experience of consciousness itself, to what Gilbert Ryle has called "the ghost in the machine". That may not be an explanation of consciousness, but it will be pretty close. (Weinberg, 1992, 45)

Chalmers (1996) considers this scenario and replies to the effect that pretty close is not good enough because it “does not explain everything that is happening in the world. To be consistent, we must acknowledge that a truly final theory needs an additional component.” (p. 126)

In my view, Chalmers is correct. If a proposed theory of everything does not explain the occurrence of experiential phenomena, the job is not done. Chalmers' argument aims to show that we don't have to wait until physicists turn in their work. We have good reasons for concluding now that they will not succeed.

Consequently, the search for a fundamental theory of conscious experience is the search for that additional component.

### §5.3.1 *Extra Ingredients vs Extraneous Ingredients*

When considering the possibility of adding elements to the scientific account of experience, I distinguish between extra ingredients and extraneous ingredients. Chalmers conflates these two categories.

In “Facing up to the Problem of Consciousness”, Chalmers (1995) surveyed the field and argued that

> … there are systematic reasons why the usual methods of cognitive science and neuroscience fail to account for conscious experience. These are simply the wrong sort of methods: nothing that they give to us can yield an explanation. To account for conscious experience, we need an extra ingredient in the explanation. (Chalmers, 1995a, 207)

Chalmers then mentions several examples of existing theories with extra ingredients such as nonlinear and chaotic dynamics, nonalgorithmic processing and quantum mechanics. Each of these is something that is already known to scientists and/or mathematicians; so, a theory of experience that incorporated one of these ingredients would not violate [EPS], the claim of Explanatory Power of Science. Such ingredients are “extra” only in the sense that they are not usually found in a reductive theory of experience.

If a theory of experience includes an ingredient not known to scientists, I would call that philosophical add-on an *extraneous* ingredient. Clearly, a theory of consciousness that succeeded only because of an extraneous ingredient would breach or deny [EPS].

In short, as I will use these terms, an *extra* ingredient is something known to scientists but which is not usually invoked by philosophical theories of consciousness; whereas, an *extraneous* ingredient is something that is unknown to scientists.

Most physicalist theories incorporate an extraneous ingredient, the causal closure principle. As originally stated, Chalmers' own theory, naturalistic dualism, incorporated causal closure. However, as we shall see, Chalmers recently shifted his position on causal closure. Nevertheless, naturalistic dualism postulates new fundamental properties unknown to physical scientists; so, it retains an extraneous ingredient.

After reviewing Chalmers' justification for naturalistic dualism, his argument against supervenience physicalism and his response to the conclusion that supervenience physicalism can not succeed, I will argue that each of these extraneous ingredients is problematic and should be abandoned.

Naturalistic dualism would still have two peculiar features, the option to take experience as itself fundamental and the double-aspect theory of information. I will then argue that a plausible theory of conscious experience can be developed from these two features of naturalistic dualism; and, that it would look very much like the theory that physicist Henry P. Stapp has been advocating for some time.

Just as removing extraneous ingredients (the causal closure and causal reduction principles) from Searle's theory of intentional action by rational agents allows us

to recognize a convergence of physics and philosophy, removing extraneous ingredients from naturalistic dualism will allow us to recognize another such convergence.

## *§5.3.2 Adding an Extraneous Ingredient*

In setting up his critique of supervenience physicalism, it seems that, Chalmers understands the dilemma: the objective is to explain experiential *phenomena*; but, supervenience is a relation between sets of *properties*.

> In general, supervenience is a relation between two sets of properties: B-properties -- intuitively, the *high-level* properties -- and A-properties, which are the more basic, *low-level* properties. … B-properties *supervene* on A-properties if no two possible situations are identical with respect to their A-properties while differing in their B-properties. (Chalmers, 1996, 33)

To get the critique rolling, Chalmers appears to assume that, for each experiential phenomenon, there is a corresponding property; namely, the property of instantiating (or exemplifying) that phenomenon. He then runs the critique with *those* properties as the B-properties which are theorized to supervene on the A-properties which are assumed (by supervenience physicalists) to be ordinary physical properties.

> A natural phenomenon is reductively explainable in terms of some lower-level properties if the property of instantiating that phenomenon is globally logically supervenient on the low-level properties in question. … If the property of exemplifying a phenomenon fails to supervene logically on some lower-level properties, then given any lower-level account of those properties, there will always be a further unanswered question: Why is this lower-level process accompanied by the phenomenon? (Chalmers, 1996, 48)

Chalmers then goes on to consider a variety of arguments (the zombie argument, the knowledge argument and others) and concludes that the phenomenal facts do not supervene on the physical facts; and, therefore, there is no hope of a reductive physical explanation of phenomenality. Physical explanations are "well suited to the explanation of structure and of *function*". However,

> Once we have explained all the physical structure in the vicinity of the brain, and we have explained how all the various brain functions are performed, there is a further sort of explanandum: consciousness itself. Why should all this structure and function give rise to experience? The story about the physical processes does not say. (Chalmers, 1996, 107)

To obtain a fundamental theory, Chalmers (1995a, 1995b, 1996) suggests taking experience itself as a fundamental feature of the world.

> There are two ways this might go. Perhaps we might take experience itself as a fundamental feature of the world, alongside space-time, spin, charge, and the like. That is certain phenomenal properties will have to be taken as basic properties. Alternatively, perhaps there is some other class of novel fundamental properties from which phenomenal properties are derived. … We could call these properties protophenomenal properties. (Chalmers, 1996. p. 126)

Taking experience itself as fundamental is identified with taking some

phenomenal (or protophenomenal) properties as fundamental. The wording suggests that we already know about phenomenal properties; protophenomenal properties would be new. In either case,

> To bring consciousness within the scope of a fundamental theory, we need to introduce new fundamental properties and laws. (1996, 126)

These new laws will supplement without contradicting any known physical law.

> Here the fundamental laws will be *psychophysical* laws, specifying how phenomenal (or protophenomenal) properties depend on physical properties. These laws will not interfere with physical laws; physical laws already form a closed system. Instead, they will be *supervenience* laws, telling us how experience arises from physical processes. We have seen that the dependence of experience on the physical cannot be derived from physical laws, so any final theory must include laws of this variety. (Chalmers, 1996, 127)

Clearly, Chalmers remains committed to the thesis of supervenience; but, it's not clear how he knows, before these laws are found, that they will be of this nature.

## *§5.3.3 Identifying Phenomenal Properties*

Oddly enough, in setting up the argument against supervenience physicalism, Chalmers does not even use the term, *phenomenal property*.

Simplifying Chalmers position by eliminating complications about local/global and logical/nomic supervenience yields a very simple initial assumption: A natural phenomenon is reductively explainable if the property of instantiating that phenomenon supervenes on lower-level properties.

After reaching the conclusion that experience can not be reductively explained by reference to physical properties, Chalmers inserts the term into the specification of what it means to take experience itself as fundamental: “certain phenomenal properties will have to be taken as basic properties”.

*To what does the term “phenomenal property” refer?*

Is an experiential phenomenon for which a reductive explanation is sought itself the phenomenal property; or, is the property of instantiating an experiential phenomenon the phenomenal property in question?

To make the question more concrete, one might naturally assume that, among all the other items that a theory of conscious experience must explain, is our old favorite, phenomenal redness. We might then ask ourselves these questions

1. Is phenomenal redness itself a phenomenal property that we could take as a basic property?
2. If not, is the property of instantiating phenomenal redness[61] a phenomenal property that we could take as a basic property?
3. If not, what else is a phenomenal property?

---

61 Or, the property of exemplifying phenomenal redness. Or, the property in virtue of which phenomenal redness occurs. Etc. I don't think much turns on which formulation is used.

If phenomenal redness and all other instances of what I have been calling experiential phenomena are actually, in Chalmers' view, phenomenal properties, I would wonder why he even mentioned “natural phenomenon” in setting up the argument against supervenience physicalism.

Since supervenience is defined as a relation between sets of properties, it would have been cleaner and simpler to avoid any reference to natural phenomena. If items such as phenomenal redness are phenomenal properties, there would be no question that they are the B-properties the analysis deals with. Once he reached his conclusion that phenomenal properties do not supervene on lower-level physical properties, Chalmers would have a natural transition to either of his alternatives, taking some phenomenal properties as basic or postulating a novel class of fundamental protophenomenal properties.

A consideration of the dialectical context of the argument suggests the same puzzle. Since Chalmers is contesting supervenience physicalism, he maximizes the impact of his arguments by conceding to that view whatever his argument does not require him to contest. Assuming that supervenience physicalists hold that phenomenal redness, if it exists at all, is a property of some kind, Chalmers' argument would have greater persuasive force if he shows that supervenience physicalism fails even granting their *propertyist* assumption about phenomenal redness.

Of course, phenomenists would object that taking phenomenal redness and its cousins as properties perpetrates a category error. If the objective is to discover how our technicolor phenomenology arises from electrochemical brain activity, it seems only natural to assume that the items of which first-person phenomenology consists are phenomena.

Taking the items of which our technicolor phenomenology consists to be phenomena, experiential phenomena in my terms, raises the question of how to engage the supervenience theorist. It's not at all clear what it means to say that one phenomenon supervenes on another.[62] Chalmers neatly avoids the question by assuming a principle of phenomenon/property correspondence.

While this principle is implicit in the initial assumption of the anti-supervenience argument, Chalmers is more explicit in a later work.

> For any distinctive kind of conscious experience, there will be a corresponding phenomenal property: in essence the property of having a conscious experience of that kind. For example, being in a hypnotic state of consciousness is a phenomenal property; having a visual experience of a horizontal line is a phenomenal property; feeling intense happiness is a phenomenal property; feeling a throbbing pain is a phenomenal property; being conscious is a phenomenal property. (Chalmers, 2010, 67)

How much distance is there between the experiential phenomenon and the corresponding phenomenal property? In the case of pain, perhaps none. Intuitively, the feeling of pain *is* the pain. Intense happiness *is* the experiencing

---

62 Suppose we have a correlation between experiential and physical phenomena, pain and C-fibers discharging, say; and, suppose further that we take the two phenomena to be non-identical. How do we know that the relation between the two is supervention or non-supervention?

of intense happiness.

Shifting to the *what it is like* idiom does not help.

> A being is conscious in the sense I am concerned with when there is something it is like to be that being. A mental state is conscious when there is something it is like to be in that state. Conscious states include states of perceptual experience, bodily sensation, mental imagery, emotional experience, occurrent thought and more. There is something it is like to see a vivid green, to feel a sharp pain, to visualize the Eiffel tower, to feel a deep regret and to think that one is late. Each of these states has a *phenomenal character*, with *phenomenal properties* (or *qualia*) that characterize what it is like to be in the state. (Chalmers, 2010, 104)

In a footnote to this passage, Chalmers explains that "In my usage, qualia are simply those properties that characterize conscious states according to what it is like to have them." (ibid.)

Here experiences are individuated or characterized according to their phenomenality. Experiencing vivid greenness is what it is like to see a vivid green. Experiencing a deep regret is what it is like to feel a deep regret. In this respect the only difference between our positions is that I would call greenness an experiential phenomenon whereas Chalmers would call it a phenomenal property.

This conclusion is further supported by Chalmers' analysis of the concept of experiential content, where he asks "what view of the content of perceptual experience is the most phenomenologically adequate?" (Chalmers, 2010, 397) His answer is illuminating.

> The view of content that most directly mirrors the phenomenology of color experience is primitivism. Phenomenologically, it seems to us as if visual experience presents simple intrinsic qualities of objects in the world, spread out over the surface of the object. When I have a phenomenally red experience of an object, the object seems to be simply, primitively, *red*. The apparent redness does not seem to be a microphysical property, or a mental property, or a disposition, or an unspecified property that plays an appropriate causal role. Rather it seems to be a simple qualitative property with a distinctive sensuous nature. (Chalmers, 2010, 398)

All of this suggests that Chalmers holds that first-person phenomenology consists of phenomenal (or qualitative) *properties* rather than experiential *phenomena,* making him a propertyist rather than a phenomenist.

In my view, this conclusion revives the puzzle concerning the nature of the "natural phenomena" referred to in Chalmers' anti-supervenience argument; but, I will set that aside rather than continue in interpretive circles. I will first examine the consequences of assuming that phenomenal redness is an example of what Chalmers calls a phenomenal property and then examine the consequences of assuming that phenomenal redness is an experiential phenomenon and that a phenomenal property is something else.

#### §5.3.3.1 An Experiential Phenomenon is a Property?

If phenomenal redness is itself a phenomenal property, we would have to take each phenomenal property as a fundamental property, which seems excessive; or,

we would have to justify saying that some phenomenal properties are more fundamental than others, which seems unlikely. In either case, it is difficult to see how the classification of phenomenal redness could enter into a psychophysical law linking experience to physical processes.

Suppose we linked experiencing phenomenal redness a particular neural firing pattern. Upon further investigation, we might discover something that we could call a psychophysical law; perhaps, something like [PPL-1]

> **[PPL-1] In trichromatic organisms like humans, whenever the structures of the brain perform the function of processing information possessing characteristics, $C_1$, $C_2$ ... $C_n$, experiencing phenomenal redness occurs to the experiencing subject associated with that brain.**

Let us assume that, by the time such a law is proposed, scientists are able to specify precisely the characteristics that the processed information must possess to play the role allotted to it in such a psychophysical law. It may eventually turn out that specific shades of phenomenal redness might be linked to specific values of the specified characteristics of information.

However fine-grained such laws turn out to be, they wouldn't require explicit mention of the classification of phenomenal redness; [PPL-1] has no deficiency compared to [PPL-1.1] or [PPL-1.2] which include appositive phrases identifying the classification of phenomenal redness.

> **[PPL-1.1] In trichromatic organisms like humans, whenever the structures of the brain perform the function of processing information possessing characteristics, $C_1$, $C_2$ ... $C_n$, experiencing phenomenal redness, a phenomenal property, occurs to the experiencing subject associated with that brain.**
>
> **[PPL-1.2] In trichromatic organisms like humans, whenever the structures of the brain perform the function of processing information possessing characteristics, $C_1$, $C_2$ ... $C_n$, experiencing phenomenal redness, an experiential phenomenon, occurs to the experiencing subject associated with that brain.**

It might be objected that [PPL-1] doesn't qualify as a fundamental law because it isn't a supervenience law or because it doesn't explain *how* experience arises from the physical processes with which it is correlated. I agree that [PPL-1] isn't a supervenience law and that it doesn't explain how experience arises; but, I deny knowing a priori that scientists will eventually discover laws of that nature.

[PPL-1] is a correlation law. It proposes a nomological relation between two sets of phenomena. By describing the circumstances in which a given experiential phenomenon occurs, it specifies *which* physical phenomena that experiential phenomenon arises from or in connection with; but, it doesn't specify how that happens.

Someone who disagreed with the constraints that Chalmers (1995b, 1996) interpolates into the proposal to take experience itself as fundamental need look no further than Chalmers (1995a) for a simpler, cleaner version of what it means

to take experience as fundamental.

> I suggest that a theory of consciousness should take experience as fundamental. ... We might add some entirely new nonphysical feature, from which experience can be derived, but it is hard to see what such a feature would be like. More likely, we will take experience itself as a fundamental feature of the world, alongside mass, charge, and space-time. If we take experience as fundamental, then we can go about the business of constructing a theory of experience.
>
> …
>
> Of course, by taking experience as fundamental, there is a sense in which this approach does not tell us why there is experience in the first place. But this is the same for any fundamental theory. Nothing in physics tells us why there is matter in the first place, but we do not count this against theories of matter. Certain features of the world need to be taken as fundamental by any scientific theory. A theory of matter can still explain all sorts of facts about matter, by showing how they are consequences of the basic laws. The same goes for a theory of experience. (Chalmers, 1995a, 210)

In this proposal, there is no mention of phenomenal or protophenomenal properties; and, there no insistence that the new laws be supervenience laws.

A theory which took experiencing as a fundamental fact about our world, need only describe and explain the interactions between experience and other fundamental entities postulated by that theory. Indeed, one could even argue that there is a contradiction inherent in Chalmers' claim that a fundamental theory would explain how consciousness arises from the brain. Something that is fundamental to the ontology of a theory is not accounted for in terms of something else.

When discussing the argument for assuming identity to explain parallel phenomenology, I offered an alternative based on Chalmers' theory of dual-aspect information: that information that is physically instantiated in the brain becomes phenomenally instantiated.

This suggestion is not detailed enough to qualify as a fundamental theory; but, psychophysical laws that specify the detail need only relate the physical and the experiential domains. One could also invoke another passage from Chalmers to theorize that the laws linking the domains will be information laws.

> What we need now is a construct to connect the domains. Information seems to be a simple and straightforward construct that is well suited for this sort of connection, and which may hold the promise of yielding a set of laws that are simple and comprehensive. If such a set of laws could be achieved, then we might truly have a fundamental theory of consciousness. (Chalmers, 1996, 287)

Whether information laws will also be supervenience laws is a matter about which philosophers are free to remain agnostic until the question is decided empirically.

* * *

Denying physical/phenomenal identity while affirming the existence of phenomenal redness yields phenomenon dualism (if one assumes the Thesis of

Phenomenism) or property dualism (if one assumes the Thesis of Propertyism). However, if there are no non-physical properties in an explanatory role, physical/phenomenal identity plus the thesis of propertyism yields nothing more than phenomenon dualism with a category error thrown in.

Chalmers is, of course, free to say that phenomenon dualism is nothing less than property dualism with a category error thrown in. Intuitively, at least one of us is perpetrating a genuine category error; and, it clearly matters which language is used. I've shown that analogous arguments have different outcomes depending on whether something like phenomenal redness is assumed to be a phenomenal property or an experiential phenomenon.

The road forks once one adds nonphysical properties in an explanatory role; meaning, that phenomenal properties are mentioned in the explanans rather than in the explanandum. At that point Chalmers and I take different forks.

### §5.3.3.2 An Experiential Phenomenon is Not a Property

Given that phenomenal redness and the other items we are trying to explain are experiential *phenomena*, scientists may then try to discover the laws that relate physical phenomena to experiential phenomena; probably aiming for laws having the form of [PPL-1], correlation laws.

It would make sense to theorize that experiential phenomena are generated by the experience generating properties of some property bearer or some property bearers. The simplest such theory would hold that experiential phenomena are explained by the experience generating properties of the brain alone. We might amend [PPL-1] as follows:

> **[PPL-1.3] In trichromatic organisms like humans, whenever the structures of the brain perform the function of processing information possessing characteristics, $C_1$, $C_2$ ... $C_n$, experiencing phenomenal redness occurs to the experiencing subject in virtue of experience generating properties of its brain which are physical$_1$ properties.**

If physical scientists eventually detect the experience generating properties of the brain, we'd have good reason to say that those properties are physical properties of the brain. We might arbitrarily say that the physical properties which are experience generating properties are *also* phenomenal properties; but, there would be no suggestion that such phenomenal properties support a claim of property dualism.

If physical scientists never detect the experience generating properties of the brain, we would probably deny the existence of experience generating properties that are physical$_1$ properties. We might then take experience as fundamental or postulate experience generating properties that are not physical$_1$ properties.

Correlation laws such as [PPL-1] may need to include a reference to nonphysical$_1$ properties not detectable by physical scientists yielding something like [PPL-1.3].

> **[PPL-1.4] In trichromatic organisms like humans, whenever the structures of the brain perform the function of processing**

> **information possessing characteristics, $C_1$, $C_2$ ... $C_n$, experiencing phenomenal redness occurs to the experiencing subject in virtue of experience generating properties of its brain which are not physical$_1$ properties.**

However, it is hard to imagine the presence of such an “in virtue of” clause increasing the explanatory power of a psychophysical law stated without it. Indeed, it is difficult to imagine how non-physical properties could play a role in a psychophysical law relating experiential phenomena to physical processes. It is even more difficult to imagine what sort of scientific experiment might selectively discriminate between [PPL-1.4] and [PPL-1.5].

> **[PPL-1.5] In trichromatic organisms like humans, whenever the structures of the brain perform the function of processing information possessing characteristics, $C_1$, $C_2$ ... $C_n$, experiencing phenomenal redness occurs to the experiencing subject in virtue of experience generating properties of an immaterial consciousness.**

Nevertheless, we may anticipate that scientists will continue to investigate the correlations between physical and experiential phenomena, thereby allowing them to refine the portion common to [PPL-1.3], [PPL-1.4] and [PPL-1.5], the portion that states a correlation law. We should also anticipate that philosophers will take up the task of explaining how the correlations occur in virtue of the experience generating properties of … something.

*How would philosophers explain the correlations scientists are expected to find?*

Such a theory would let us assume that Chalmers' argument is still effective to show that physical properties will never explain more than structure and function. We would then conclude that experience generating properties could not be physical$_1$ properties.

### §5.3.3.3 Rejecting Properties Unknown to Scientists

In one scene in Moliere's play, *The Hypochondriac*, a student is taking a final exam. When asked to explain why opium makes people sleepy he replies to the effect that opium promotes sleep because it has a sleep promoting property.

To many, that student postulating that property is the archetypal example of pointless intellectual effort. It is difficult to remember that the satirical targets of the play were the student who offered and the professors who accepted the claim that opium has a dormitive property *as a finished explanation* (rather than, say, as a research proposal in search of funding). Scientists who actually went looking for the sleep promoting property of opium might have found what scientists did in fact eventually find, that the shape of opiate molecules are such that they can dock at the natural endorphin receptor sites in the brain. That may not be the full story of the sleep promoting property of opium; but, surely, it is part of the story.

Similarly, I see nothing wrong with postulating that experience is generated by the experience generating properties of the brain; ***provided***, *that scientists actually go looking for those properties*.

The problem is that both of the scenarios under consideration present us with the situation in which the search has ended but nothing was found.

In one scenario, physicists have already produced a theory of everything that doesn't explain experiential phenomena. In such a case, postulating properties unknown to physicists to explain experiential phenomena would be an act of desperation, one whose only virtue is that it would allow philosophers to claim success after scientists admitted failure.

In the other scenario, where philosophers have found good reasons to believe that no physical theory will explain more than structure and function, it is difficult to see how postulating properties unknown to scientists helps in the slightest.

I may be somewhat eccentric; but, I would prefer to see a *scientific* theory that explains physical and experiential phenomena. However, the epistemological journey is not like a relay race in which scientists carry the baton as far as they can before passing it off to philosophers who then assume their way across the finish line. Consequently, I reject the option to postulate fundamental properties of matter extraneous to the scientific account of the physical universe.

It may be that no scientific theory will ever explain how experiential phenomena arises from physical$_1$ phenomena. In that case, adopting mysterianism seems more attractive to me than assuming extraneous ingredients to obtain the appearance of an explanation.

### *§5.3.4 And The Job is Still Not Done*

Chalmers' (1996) book is a slightly altered version of his dissertation. Both were titled The Conscious Mind; but, the book was subtitled “In Search of a Fundamental Theory” whereas the dissertation was subtitled “In Search of a Theory of Conscious Experience”.

Together, the subtitles are revealing. Chalmers is certainly looking for a fundamental theory; but, it's a theory of consciousness qua experience only. He does not discuss the experiencing subject. So, even if one could work out the details of the theory with respect to phenomenal properties and ended up with a theory of experience, there is still a problem: *the job is still not done*.

One way to engage this problem is to consider how we are to understand the key claim of Chalmers' justification for property dualism: that something – some *one* thing, the brain – instantiates the property of instantiating a phenomenon.

I understand instantiating a property as *having* that property; so, I deny that my brain instantiates the experiential phenomena that I experience, that I *have*. As the experiencing I of all the experiential phenomena that I experience, I would say that *I* have or instantiate the phenomena that I experience. This is the essential “mineness” of experiencing.

Do I also instantiate the property of instantiating some specified phenomenon? If so, I am a property bearer.

Now, I've already concluded that I, this experiencing subject, exist and that I am

not identical to my brain; so, concluding that I am a property bearer would have a serious consequence. Given the meaning of “substance” as property bearer, it would follow that I am a substance distinct from my brain; and, substance dualism would follow.

To avoid deriving substance dualism so easily, I deny instantiating the property of instantiating experiential phenomena. Instead, I might (at least initially), hypothesize that my brain instantiates any properties in virtue of which I experience experiential phenomena.

However, as soon as I address the phenomenology of willing, another dilemma arises. Unless I have the property of being able to initiate intentional actions, I am an epiphenomenon who denies being an epiphenomenon; but, if I assert having the property of being able to initiate intentional actions, I will have affirmed interactive substance dualism.

Given that one rejects both eliminative materialism and identity theory physicalism as to the experiencing subject, there appears to be no way to deny epiphenomenalism without affirming substance dualism (and vice versa).

Chalmers hasn't dealt with these issues explicitly; but, given the recent shift in his position on the causal closure of the physical, it will be difficult for him to maintain that the experiencing subject is identical to its brain.

#### §5.3.4.1 Rejecting Causal Closure

As originally stated, premise 4 of Chalmers' argument for naturalistic dualism is the causal closure principle: “The physical domain is causally closed”. (Chalmers, 1996, 161)

Chalmers admitted that epiphenomenalism is a possible consequence of his view.

> A problem with the view I have advocated is that if consciousness is merely naturally supervenient on the physical, then it seems to lack causal efficacy. (Chalmers, 1996, 150)

In response, Chalmers denied that his view entails epiphenomenalism *in a strong sense*; but, he acknowledged that his view “implies a weak form of epiphenomenalism, and it may end up leading to a stronger sort”. (1996, 160)

In The Character of Consciousness, Chalmers acknowledged that he was too quick to reject interactionism; and, put interactionism, epiphenomenalism and neutral or Russellian monism on a more or less equal footing.

> ... the proper conclusion of the anti-materialist arguments is disjunctive. The choice among the three disjuncts rests on further and largely independent considerations. As things stand, the choice is wide open. At the end of the day, the choice will come down to which of the disjuncts yields the most successful detailed theory, in light of a well-developed science of consciousness. (Chalmers, 2010, xix)

In this work, Chalmers wrote more sympathetically toward interactionism, including the interpretation of quantum mechanics that provides a causal role for consciousness in the collapse of the wave function. However, until more recently, he did not explicitly reject the causal closure principle.

At a recent workshop in Gottingen, Germany, Chalmers acknowledged that philosophers often reject dualism because they assume that physics is causally closed, “leaving no role for consciousness”. However, he then argues that “In fact, physics leaves a giant causal opening in the collapse process”, a gap “perfectly suited for consciousness to fill”. (Chalmers, 2015, @49:00)

This approach removes an extraneous ingredient (the causal closure principle) commonly found within physicalist accounts of consciousness. In my view, that's a point in favor of Chalmers' new approach.

That said, it seems clear that Chalmers now faces a difficult choice: embrace substance dualism or distinguish his views from it. If consciousness has a causal role, it seems reasonable to say that it has a property; although, it may not be clear how we should speak about that property. Should we speak about the property of being able to play that causal role; should we speak about some property (or properties) in virtue of which consciousness is able to play that causal role; or, should we speak in some other manner?

In any case, there are consequences to giving consciousness a causal role in wavefunction collapse.

First, one is committed to denying brain/consciousness identity. The brain, as a material object, is subject to the Schrodinger equation; so, instead of collapsing the wave function of a quantum system with which it came into contact, it would become entangled with it. Consequently, if consciousness has a role in collapsing the wave function, consciousness is distinct from its brain.

Secondly, one forfeits the possibility of finding a reductive explanation. Chalmers explains:

> One special attraction of quantum theories is the fact that on some interpretations of quantum mechanics, consciousness plays an active role in "collapsing" the quantum wave function. Such interpretations are controversial, but in any case they offer no hope of explaining consciousness in terms of quantum processes. Rather, these theories assume the existence of consciousness, and use it in the explanation of quantum processes. At best, these theories tell us something about a physical role that consciousness may play. They tell us nothing about how it arises. (Chalmers, 1995a, 208, fn. 2)

*How does this help construct a fundamental theory of consciousness?*

Once one admits the existence of a property bearing consciousness (consciousness qua subject or qua agent) to explain intentional action and/or the collapse of the wave function, that consciousness may also be invoked in the attempt to explain consciousness qua experience. In such a case, consciousness (qua subject) would not be an extraneous ingredient in a theory of consciousness (qua experience).

### §5.3.4.2 The Updated Alternative

As an alternative to taking phenomenal or protophenomenal properties to be fundamental properties, the solution proposed here is that experiential phenomena occurs when information that is/was physically instantiated in the

brain becomes phenomenally instantiated in the experience of a subject not identical to the brain with which it is associated. I call this the Hypothesis of Phenomenally Instantiated Information, HPII.

How information is instantiated phenomenally remains a mystery; but, we have good reasons to believe *that* information which is physically instantiated while en route to the brain becomes phenomenally instantiated in experience … somehow.

Now, we no longer need to postulate non-physical properties of the brain because we can get by with properties that explain structure and function. We need to postulate that the brain has properties that explain its function of processing information; but, these would be physical$_1$ properties. We would then need to say that there is some event in which physically instantiated information becomes phenomenally instantiated – that there is a hand-off, so to speak, from brain to consciousness (qua experiencing subject).

Such a proposal invites the speculation that the experience generating properties of the brain somehow cause (or mediate or facilitate, etc.) the transformation of information from being physically instantiated to being phenomenally instantiated.

This vast potential for further speculation doesn't necessarily mean that HPII is not a fundamental theory; but, there is a risk that it is only fundamental in the sense that it may be as good as it gets. We will have to wait for further bulletins from the frontiers of science.

### §5.3.4.3 Rejecting Supervenience

John von Neumann is said to have been unhappy with the ad hoc nature of the quantum/classical divide assumed by the Copenhagen interpretation. In his analysis of the measurement problem, von Neumann showed that one could move the 'cut' – the dividing line between the part of the world that is described quantum mechanically and the part of the world that is described classically. One could even put the entire body and brain of the experimenter and the whole physical universe in the quantum mechanically described portion of the world, subjecting the entire physical universe to the jurisdiction of the Schrodinger equation. Doing so does not change the result of the calculation; but, it does require something not subject to the Schrodinger equation – something “outside the calculation” according to von Neumann – to collapse the wavefunction.

In von Neumann's view, it was the 'abstracktes Ich' – the abstract I – of the experimenter that caused the collapse of the wavefunction; but, he did not elaborate.

The abstract I has generally been taken to refer to the mind or consciousness of the experimenter. Early supporters of the von Neumann interpretation, London and Bauer, stated the implication clearly. “We note the essential role played by the consciousness of the observer in this transition from the mixture to the pure case.” (1939, 251)

The von Neumann interpretation was further developed by Eugene P. Wigner; and, in recognition of his contribution, the theory is often known as the von

Neumann/Wigner interpretation or, more popularly, as the Consciousness Causes Collapse theory.[63]

> It was not possible to formulate the laws of quantum mechanics in a fully consistent way without reference to the consciousness. … even though the dividing line between the observer, whose consciousness is being affected, and the observed physical object can be shifted towards the one or the other to a considerable degree, it can not be eliminated. … the very study of the external world led to the conclusion that the content of the consciousness is an ultimate reality. (Wigner, 1962, 169)

In a later paper, Wigner went draw the conclusion that threatens non-dualistic forms of physicalism.

> the regularities which modern physical theory – that is quantum mechanics – furnishes are probability connections between subsequent observations, i.e. contents of the consciousness of the observer. The primitive facts in terms of which the laws are formulated are not positions of atoms but the result of observations. It seems inconsistent, therefore, to explain the state of mind of the observer, his apperception of the result of an observation, in terms of concepts, such as positions of atoms, which have to be explained, then, in terms of the content of consciousness. (Wigner, 1969, 96)

We don't have to wait for the scenario that Chalmers considered, where physicists come up with a Theory of Everything physical that doesn't explain the occurrence of experiential phenomena, to draw the conclusion that most physicalist theories of brain/consciousness relations are false. If Wigner was right there is no hope for supervenience physicalism, which term I will here take to include any theory that the brain/experience relation is one of supervenience.

Now, it may be objected, that the von Neumann/Wigner interpretation is a minority view among physicists; and, of course, that's true; but, its unpopularity is not a valid argument in favor of any form of physicalism incompatible with it.

More to the point, the von Neumann/Wigner interpretation is an interpretation of quantum mechanics that has not yet been falsified by empirical research; consequently, advocates of supervenience physicalism, reductive physicalism and other forms of physicalism incompatible with the von Neumann/Wigner interpretation effectively presuppose its falsity.

I have no objection to philosophers advocating forms of physicalism incompatible with the von Neumann/Wigner interpretation. My objection is that such a philosophy – even assuming that arguments on its behalf survive philosophical objections – can only be accepted provisionally, the provision being the assumption that the von Neumann/Wigner interpretation is eventually falsified.

Certainly, reductive physicalists may say something similar about a philosophy of consciousness that presupposes that the von Neumann/Wigner interpretation is

---

63 The von Neumann/Wigner interpretation is also be known as the subjective reduction theory; particularly when a contrast with an objective reduction theory such as the Penrose-Hameroff theory of orchestrated objective reduction is intended. However, among physicists, “subjective interpretation” can also refer to an instrumental approach rather than an ontological approach to interpreting the formalism of QM. The von Neumann/Wigner interpretation is not subjective in that sense.

true. No such theory can be accepted except provisionally. Nevertheless, consistency with the von Neumann/Wigner interpretation may be the basis for classifying a philosophy of consciousness as a form of physicalism despite also being a form of dualism. It would be a dualistic form of physicalism.

In any case, the point is that the claim of supervenience is not a finding of physics. It may someday become a finding of physics; provided, that the von Neumann/Wigner interpretation is empirically falsified some sweet day in the future. Until then one may dispense with the claim of supervenience, just as one may dispense with the claim of causal closure, without fear of becoming inconsistent with the current state of empirical research.

### §5.3.4.4 Assessment of Naturalistic Dualism

Overall, I agree with Chalmers' rejection of supervenience physicalism; but, I disagree with his preferred alternative, a theory I take to be a property dualistic supervenience theory. Chalmers speaks about taking experience as fundamental; but, assuming the thesis of supervenience makes experience dependent on something physical; and, that makes experience non-fundamental, at least in my view.

I prefer the pristine clarity of the proposal to take experience as fundamental found in Chalmers (1995a) over the version in Chalmers (1995b and 1996) which imposes constraints on what it means to take experience as fundamental. Eliminating the ingredients of the constrained version of naturalistic dualism that are extraneous to the physicists account of the physical universe – properties unknown to scientists, causal closure and supervenience – takes us back to the pristine version.

After clearing away the deadwood, one is left to wonder whether a plausible theory of the brain/experience relation can be constructed by taking experience itself as fundamental and incorporating a version of Chalmers' dual-aspect theory of information. I believe it can be and I've used a version of Chalmers' dual-aspect theory of information as an alternate way of explaining the physical/phenomenal isomorphism that prompts some physicalists (e.g. Clark, 1994) to assume physical/phenomenal identity.

The Hypothesis of Phenomenally Instantiated Information assumes that information that is physically instantiated in the brain becomes phenomenally instantiated. Naturally, one may now speculate about the nature of the hand-off event; and, suspicion naturally falls on an event that physicists have defined in terms of information transfer, quantum wavefunction collapse.

## §6 *The Physics of Consciousness is Quantum Physics*

> The Copenhagen interpretation describes what happens when an observer makes a measurement, but the observer and the act of measurement are themselves treated classically. This is surely wrong: Physicists and their apparatus must be governed by the same quantum mechanical rules that govern everything else in the universe. (Weinberg, 2005, 32-33)

The nature of the link, if any, between consciousness and quantum mechanics turns on the resolution of the so-called measurement problem – getting right what the Copenhagen interpretation gets wrong, the quantum/classical divide, the bifurcation of the world.

Assuming a quantum/classical divide allows one to assume that the wavefunction describing a quantum system collapses on contact with the measuring system, a classical object, avoiding the measurement problem because the quantum system is separated from the observing physicist by the measuring instrument.

On the other hand, if the measuring instrument is subject to the Schrodinger equation, it would become entangled with the quantum system; and, therefore, unable to collapse a superposition of all possible values down to the single value actually observed. The problem only gets worse once the body and brain of the physicist is included in the measuring system and becomes entangled with the quantum system under observation.

> In other words, if the whole physical universe were composed only of microphysical entities, as it should be according to the atomic theory, it would be a universe of evolving potentialities ... but not of real events. (Jammer, 1974, 474)

*How do we explain the occurrence of actual events?*

At one time, one might have argued that the mathematical formalism should be interpreted instrumentally, as merely a means to calculate the probabilities associated with the various possible outcomes of a measurement. However, as mentioned previously, recent no-go theorems support the conclusion that the wavefunction should be interpreted ontologically, as a description of the physical reality whose behavior it predicts. Doing so makes the collapse a genuine phenomenon, an actual physical event, rather than an indicator of the change in our knowledge of the quantum system under observation.

Presumably, if the collapse of the wavefunction is an actual physical event, it has a cause. But, if everything material would become entangled with the quantum system under observation, something “outside of the calculation” (von Neumann, 1955, 421) must be found to collapse the wavefunction. This has resulted in a search for what collapses the wavefunction. There are many such theories; but, not all of them attempt to relate some aspect of consciousness to quantum phenomena.

In von Neumann's own interpretation, the actual observer, the *abstracktes Ich* – the abstract I – of the experimenter, is responsible for the collapse. Taking the abstract I as an immaterial consciousness is the origin of the Consciousness Causes Collapse perspective. In §6.4, I will consider a modern version of the von

Neumann interpretation, Stapp's Dual-Aspect Reduction Event theory, DARE.

Some physicists prefer not to grant consciousness the role it has in the von Neumann interpretation; and, many theories have been offered that postulate an objective reduction of the wavefunction; but, not all such theories make an attempt to explain subjective experience. In §6.3, I'll consider a version that does: the theory of Orchestrated Objective Reduction, Orch OR, proposed by Roger Penrose and Stuart Hameroff.

In §6.2, I'll consider the Spin-Mediated Consciousness Theory, SMCT, proposed by Huping Hu and Maoxin Wu. SMCT is unusual in that it presupposes that the human body is electromagnetic as well as biochemical. As the name of the theory suggests, the quantum phenomenon that does the work is particle spin. SMCT has the potential to explain human sensitivity to ambient magnetic fields and some parts of it are testable with current technology.

Each of these theories is an example of what I will call a quantum consciousness theory; meaning, a theory that relates consciousness to the quantum brain, the brain described quantum mechanically. However, before beginning this inquiry, I want to relate all such theories to the Warm, Wet Brain Objection and its offshoot, the Bird Wing Objection.

## §6.1 No Go for Quantum Phenomena in the Brain?

It was once thought that the warm, wet, noisy brain would be hostile to quantum phenomena because quantum coherent states would not last long enough to accomplish anything of biological significance; but, this position is hotly contested.

In a criticism directed mainly against the Orch OR theory, Tegmark (2000) calculated the decoherence rates for microtubular processes and concluded that “there is nothing fundamentally wrong with the current classical approach to neural network simulations” because a quantum superposition would decay in a fraction of the time required for a neuronal process to run its course.

In a direct reply, Hagan et al. (2002) argued that Tegmark had based his calculations on an incorrect set of assumptions; and, that calculations based on the assumptions that Orch OR actually made showed that coherent states would survive long enough to be neurophysiologically relevant. Two other considerations were offered as well. It was argued that there may be natural processes of quantum error correction; and, it was proposed that there may be natural processes that protected a quantum coherent state from environmentally induced decoherence.

There has been no further development of the first suggestion; and, some would argue that the difficulty of engineering quantum error correction routines for technological quantum computers makes it unlikely that nature solved the problem eons ago. See Litt et  al. (2006) for a summary.

The second suggestion, however, was adopted and made more specific by Hameroff et al. (2002). There it was argued on the basis of medical evidence that the crucial process was electron mobility in hydrophobic (water-excluding)

pockets in proteins in the brain's microtubular structure. Due to inherent uncertainty as to the location of an electron, a process dependent on electron mobility – for example, alterations to the conformation state of various brain proteins – would be a quantum phenomenon.

For other theoretical arguments to the effect that environmental factors could preserve rather than destroy quantum coherent states, see Hartmann et al. (2006) and Li and Paraoanu (2009).

There are a number of other possible replies to this objection as well.

One might try to evade the argument by noting that not all quantum phenomena are equally vulnerable to environmental decoherence. “Nuclear spins are so weakly coupled to the environmental degrees of freedom that, under some circumstances, phase coherence times of five minutes or perhaps longer are possible” (Fisher, 2015, 594 (citations omitted)). Consequently, a theory such as SMCT that relied on quantum spin would be less vulnerable to the warm, wet brain objection.

Similarly, Stapp's Dual-Aspect Reduction Event theory relies, in part, on the Quantum  Zeno Effect, QZE, a well documented resistance to environmental influences in the case of rapidly repeated measurements.

The most decisive reply to no-go arguments based on decoherence time is simply the increasing evidence that quantum phenomena do, in fact, occur in biological environments and play significant neurophysiological roles. For examples, in recent years, the quantum phenomena have been invoked to help explain avian magnetoreception (Kominis, 2008), photosynthesis (O’Reilly and Olaya-Castro, 2014) and the sense of smell (Bittner et al., 2012).

In view of all of this, it seems clear that, despite the warm, wet brain objection, we have defensible reasons for believing that the brain must be described quantum mechanically just to describe neural phenomena accurately; so, one may plausibly argue that a successful theory of brain/consciousness relations must relate consciousness (qua experience and/or qua subject of experience) to the quantum brain.

Less clear is whether a successful theory of brain/consciousness relations must invoke quantum phenomena *for explanatory purposes*.

In principle, one might argue that quantum phenomena in the brain only perform biologically useful functions unrelated to the occurrence of consciousness. For example, Litt et  al. (2006) argue that a theory explaining how birds fly could rely on the structural properties of the bird's wing without mentioning the quantum phenomena involved in bonding together molecules of keratin, the dominant ingredient of bird wings.

> The relevance of atomic bonding properties to the structure of wings does not necessitate their involvement in explaining flight, because aerodynamic mechanisms have proven sufficiently powerful to explain the phenomenon. Only if specific, flight-relevant geometric or tensile features arose purely from atomic bonding properties in feathers would it make sense to import these details into our explanations of bird flight. Because no such special properties are found in existing examples of wings, atomic bonding is not

> relevant to explaining bird flight. Similarly, there appear to be no special quantum mechanical properties needed to explain psychological and neurological phenomena. (Litt et al., 2006, 600-601)

However, there is mounting scientific evidence that there *may* be a need to refer to the details of low level quantum phenomena to explain higher level psychological phenomena.

Research in the field of quantum cognition has revealed what appear to be interference effects in human decision making (Pothos and Busemeyer, 2009; Busemeyer et al., 2011; Pothos and Busemeyer, 2013), perception of ambiguous figures (Conte et al., 2009; Manousakis, 2009) and the effect of question ordering (Wang et al., 2014)

In such cases the outcomes of psychology experiments can be better predicted using the same math that is used to describe quantum events. The simplest theory to explain such results is that low level quantum phenomena affect higher level cognitive process; but, until scientists examine the brain from a quantum mechanical point of view, we will not know whether this is so.

* * *

Given that we have defensible reasons for thinking that consciousness is related to the quantum brain, we may turn our attention to the theories themselves. Due to the ontic breakthrough in the philosophy of physicists themselves, I am only considering theories that are realistic about wavefunction collapse.

## §6.2 Spin-Mediated Consciousness Theory

Hu and Wu cite developments in mathematical physics to argue that particle spin is a fundamental, self-referential process – a process possibly more fundamental than spacetime – and conclude that "... spin is the process driving quantum effects" (2004a, 634). On this basis they advance the first postulate of SMCT: that "consciousness is intrinsically connected to quantum spin" (2004a, 636). Hence, the slogan: *spin is the mind-pixel*.

The second postulate of SMCT specifies which spins are the mind-pixels

> ... the mind-pixels of the brain are comprised of the nuclear spins distributed in the neural membranes and proteins, the pixel-activating agents are comprised of biologically available paramagnetic species such as $O_2$ and NO ... (Hu and Wu, 2004a, 636).

Both $O_2$ and NO (nitric oxide, a neurotransmitter) have unpaired electrons and both are paramagnetic; meaning, that they are attracted to an externally applied magnetic field. This led Hu and Wu to speculate that "$O_2$ and NO may serve as spin-catalysts in consciousness-related neural biochemical reactions" (2004a, 634).[64]

---

64 Hu and Wu also note that $O_2$ and NO have unpaired electrons which may make electron spin significant; and, in later work, these researchers effectively modify the second postulate to read “nuclear and/or electronic spin”. For example, they say “the nuclear/electronic spins are

The idea seems to be that paramagnetic molecules are impacted by ambient magnetic fields and generate a relatively strong magnetic field which affect the nuclear spin states of certain nuclei, $^{1}H$, $^{13}C$ and $^{31}P$. There follows a two way interaction, the third postulate of SCMT.

> ... action potential modulations of nuclear spin interactions input information to the mind-pixels and spin chemistry is the output circuit to classical neural activities ... (2004a, 636).

Hu and Wu have conducted some intriguing research showing that magnetic pulses can convey information to the brain.

> We found that applying magnetic pulses to the brain when an anesthetic was placed in between caused the brain to feel the effect of said anesthetic as if the test subject had actually inhaled the same. We further found that drinking water exposed to magnetic pulses, laser light or microwave when an anesthetic was placed in between also causes brain effects in various degrees. (Hu and Wu, 2006b, 291)

### *§6.2.1 Human Electromagnetic Sensitivity*

There is a large body of evidence suggesting that human beings are sensitive to electromagnetic fields; and, some have argued that external electromagnetic fields can contribute to various kinds of anomalous experiences. For example, Michael A. Persinger has argued that tectonic strain generates electromagnetic fields to which humans are sensitive and that such fields could explain some reported UFO sightings (1976, 1981, 1982, 1983).[65]

Similar research also supports a link between ambient EM fields and spontaneous psi events (1979), telepathic dreams (Persinger and Krippner, 1985), poltergeist activity (Gearhart and Persinger, 1986) and bereavement hallucinations (Persinger, 1988). However, even if all these correlations hold up under scrutiny, none of them reveal the means by which the nervous system interacts with ambient electromagnetic fields.

Even the induction of mystical experiences in the laboratory via electromagnetic stimulation (Persinger et al., 2010; Saroka et al., 2010) does not reveal the mechanism by which these effects take place. Only if it can be shown that some or all of these effects result from interactions between the external electromagnetic stimulus and endogenous nuclear and/or electronic spin would the SMCT gain support from them. At present there is only circumstantial experimental evidence showing that this may be the case; for example, (Persinger et al., 2013).

Nevertheless, SMCT offers a credible mechanism to explain the effect of ambient electromagnetic fields on the brain; so, it has the virtue of being able to explain a large body of evidence.

---

proposed to be the mind-pixels which interact with the brain through quantum effects, modulating and being modulated by various classical brain activities such as the action potentials" (Hu and Wu, 2008, 26).

65 For a criticism of the Tectonic Strain Theory, TST, see Rutkowski (1984). See also Persinger's (1985) rebuttal and Rutkowski's (1986) rejoinder.

## *§6.2.2 Avian Magnetoreception*

Many life forms are sensitive to the Earth's magnetic field. Birds, for example, appear to navigate by orienting themselves relative to the Earth's magnetic fields. How various organisms sense the geomagnetic field is the subject of considerable research. Competing theories rely on a either a magnetite-based mechanism or radical pair mechanism.

According to the radical pair theory,

> … the avian compass relies on magnetically sensitive radical pairs formed by photoinduced electron transfer reactions. The cryptochromes in the retina of migratory birds provides a potential physiological implementation of such a mechanism. (Cai and Plenio, 2013, 1)

Each radical ion has an unpaired electron and these electrons, in each pair of radical ions formed in a reaction with a magnetic photon; and, researchers wondered whether these electrons had to be quantum entangled to make the radical-pair mechanism, RPM, work.

> We find that the answer largely depends on the radical-pair lifetime. For specific realizations of the RPM, e.g., those in recent spin chemistry experiments, entanglement features prominently and can even serve as a signature of the underlying spin dynamics. However, when the radical-pair lifetime is extremely long, as is believed to be the case in the molecular candidate for magnetoreception in European robins, entanglement does not seem to be significant. (Cai et al, 2010, 1)

A short coherence lifetime is problematic for the radical-pair mechanism. The spin-coherence must last long enough to accomplish something of biological significance.

> Critical to this mechanism is the long lifetime of the radical-pair spin coherence, so that the weak geomagnetic field will have a chance to signal its presence. It is here shown that a fundamental quantum phenomenon, the quantum Zeno effect, is at the basis of the radical-ion-pair magnetoreception mechanism. The quantum Zeno effect naturally leads to long spin coherence lifetimes, without any constraints on the systems physical parameters, ensuring the robustness of this sensory mechanism. (Kominis, 2008)

According to Kominis, sensitive human constructed magnetometers rely on the QZE; so, invoking the QZE to explain how the radical-pairs last long enough to participate in a quantum measurement seems plausible. But that still leaves the question of how an organism's sensitivity to magnetic fields is translated into information that would allow the organism to orient itself relative to the geomagnetic field.

It seems that the alignment between the geomagnetic field and the radical pairs alters the results of recombining the pieces of the pairs. (Ritz et al., 2000) Thus, information about bird's orientation relative to the geomagnetic field is encoded in the radical pair recombination results.

These researchers further speculate that the information obtained about the geomagnetic field is translated "into a modulation of visual perception" (p. 708) and that studying the effects of magnetic fields on the behavior of the organism may reveal how this translation occurs.

Other researchers have been more explicit in stating that the radical-pair mechanism

> … postulates a close link to vision and supposes that the animals can see the position of the geomagnetic North as a visible pattern superimposed on the picture of the environment. (Valkova and Vacha, 2012, 461)

* * *

As discussed above, Hu and Wu proposed a two way interaction between consciousness and world via nuclear/electronic spin. Spin chemistry is supposed to be the *output* process, “spin chemistry can serve as the bridge to the classical neural activities since biochemical reactions mediated by free-radicals are very sensitive to small changes of magnetic energies” (Hu and Wu, 2006a, 636-637).

However, in the story of avian magnetoreception as told by the radical pair theorists, the radical pair mechanism seems more like a part of the input process. The radical-ion-pair detects the incoming photon, information is extracted and sent to the brain.

Nevertheless, the reliance on spin chemistry for an important biological process makes it reasonable to suspect that spin chemistry may be applicable elsewhere, a conclusion that is generally supportive of SMCT.

## *§6.2.4 The Quantum Entangled Brain*

In early work, Hu and Wu favored the idea that nuclear spin states become quantum entangled. They speak of nuclear spin ensembles or networks and say that "consciousness emerges from the collapses of those entangled quantum states which are able to survive decoherence ... " (2004a, 636).

As used here "consciousness" seems to mean "experience". In considering alternatives should there be no large scale quantum coherence of neural spin networks, Hu and Wu propose a dualistic approach in which

> ... mind has its own independent existence and reside in a pre-space-time domain. Then, the question becomes how does mind process and harness the information from the brain so that it can have conscious experience? (2004a,640)

In later work, however, Hu and Wu seem to have adopted this dualistic perspective without abandoning the postulate of large scale quantum coherence. In Hu and Wu (2010), they speak about human consciousness as a limited or individuated version of a universal Consciousness (capitalized) which has both a transcendental aspect and an immanent aspect. The transcendent aspect of Consciousness interacts with physical reality through self-referential spin. The individuated human consciousness (lower-case) also has a transcendent aspect (responsible for free will) and an immanent aspect (supports awareness of our immediate experience); presumably, these aspects of human consciousness also interact with the brain by affecting particle spin.

## *§6.2.5 A Recent Refinement of SMCT*

In a recent refinement of SMCT, Matthew P. A. Fisher argued that the spin of an atomic nucleus could be used for quantum computation only if that nucleus had a spin of ½. A nucleus with a 0 spin would not interact with magnetic fields and a nucleus with a spin > ½ would decohere too quickly.

> Magnetic and electric field perturbations cause quantum decoherence of the nuclear spin – anathema to quantum processing – so that the "coherence time", $t_{coh}$, must be maximized when seeking a possible biological arena for nuclear spin processing.

In the biochemical setting electric fields are the primary source of decoherence for nuclei with I > ½, while I = ½ spins are more weakly decohered only by magnetic fields. ... Thus, the element hosting the putative neural qubit must have a nuclear-spin of I = ½.

> Among the most common biochemical elements, carbon, hydrogen, nitrogen, oxygen, phosphorus and sulfur, and the common ions $Na^+$, $K^+$, $Cl^-$, $Mg_2^+$ and $Ca_2^+$, besides hydrogen, only phosphorus has a nucleus with spin I = ½. *This identifies the phosphorus nucleus as our putative neural qubit*. (Fisher, 2015, 594)

Fisher then provides an elaborate theory identifying the phosphate ion as the "qubit transporter" and the so-called Posner molecule, $Ca_9(PO_4)_6$ as the "qubit memory".

The hydrolysis of ATP to AMP produces pyrophosphate ions, $P_2O_7$. A pyrophosphate ion is then converted into 2 inorganic phosphate in an enzyme catalyzed reaction from which, Fisher theorizes, the two phosphate ions emerge quantum entangled. If the two phosphate ions are then incorporated into different Posner molecules, those molecules will be entangled as well.

Fisher then concludes

> The chemical binding of multiple Posner molecules with entangled nuclear spins might allow for complex quantum processing. ... Clouds of multiple entangled Posner molecules can induce correlated, non-local binding reactions, a powerful setting for quantum processing. (Fisher, 2015, 599)

One would expect that a theory this detailed would be testable.

> If the phosphorus nuclear spins inside Posner molecules are playing a functional role in the brains of mammals (or, possibly, other vertebrates), then perturbations of the nuclear spins might have behavioral manifestations. (Fisher, 2015, 601)

Fisher speculates

> If two lithium ions can be incorporated inside the Posner molecules during molecule formation (replacing the central divalent calcium cation) they would tend to decohere the phosphorus nuclear spins, offering a possible explanation for the remarkable efficacy of lithium in tempering mania in patients with bipolar disorder. If this is indeed the mechanism, one might expect a lithium isotope dependence on the behavioral response. Remarkably, a lithium isotope dependence on the mothering behavior of rats chronically fed either $^6Li$ or $^7Li$ – having elevated or depressed alertness levels, respectively – has indeed been reported. (Fisher, 2015, 601)

Fisher is referring to the work of Sechzer et al. (1986); but, those authors did not determine the mechanism by which isotopically pure lithium

compounds could have differential behavioral effects. They did, however, point out that there is a large mass difference, 17%, between $^{6}$Li and $^{7}$Li, due to the extra neutron in $^{7}$Li.

This extra neutron would also give $^{7}$Li a spin of I > ½ and Fisher is reported to believe that this spin difference accounts for the difference in behavioral effects because the difference in mass would have little impact in the watery environment of the brain. (Ouellette, 2016)

Obviously, more research will be needed to identify the reason for the differences in the effects of isotopically pure $^{6}$Li and $^{7}$Li compounds.

### *§6.2.6 Assessment of SMCT*

Spin Mediated Consciousness Theory shows great promise as an explanation for a variety of puzzling phenomena; but, even if it turns out that the neural correlate of consciousness qua experience involves quantum spin, we still have no clue as to *how* experience arises. We would only know that certain experiential phenomena occur while certain physical phenomena involving quantum spin are occurring. Even if we knew that the physical phenomenon counts as processing information from the external world conveyed to the brain via ambient electromagnetic fields, we would not thereby know how experiential phenomena arises from quantum information processing in the brain.

Perhaps, once the neural correlate is known, the arising of experiential phenomena in relation to that correlated phenomena must be accepted as a brute fact.

## §6.3 Orchestrated Objective Reduction, Orch OR

> The 'hard problem' of incorporating the phenomenon of consciousness into a scientific world-view involves finding scientific explanations of qualia, or the subjective experience of mental states. On this, reductionist science is still at sea. Why do we have an inner life, and what exactly is it? … Our viewpoint is to regard experiential phenomena as also inseparable from the physical universe, and in fact to be deeply connected with the very laws which govern the physical universe. (Hameroff and Penrose, 1996, 37)

In the Penrose/Hameroff theory of Orchestrated Objective Reduction of the quantum wavefunction, Orch OR, quantum information is preserved in a coherent (superpositioned) state by natural, biological means of isolating the quantum system from its environment until it reaches the point in the brain where the superposition collapses due to gravitational influences, thereby performing a quantum computation resulting in a moment of conscious experience.

### *§6.3.1 Shifting to An Identity Claim*

The specifics of OOR have changed as the theory has evolved over the years. Early versions of OOR were consistent with phenomenon dualism. Hameroff and Penrose (1996, 2) wrote that “Sequences of OR events give rise to a 'stream' of

consciousness". That is most naturally taken to imply that a moment of conscious experience is not identical to the reduction event with which it is associated. A moment of conscious experiencing would not give rise to itself.

In contrast, the latest version of OOR adopts an identity theory. In criticizing a thought experiment proposed by Koch and Hepp (2006), Hameroff and Penrose (2014) wrote that they appear to have assumed that

> Orch OR followed the version of the Copenhagen interpretation in which conscious observation, in effect, *causes* quantum state reduction (placing consciousness outside science). This is precisely the opposite of Orch OR in which consciousness *is* the orchestrated quantum state reduction given by OR. (p. 67)

When I first read this passage, I was so startled by what appeared to be a shift in position not singled out for explicit commentary that I emailed corresponding author Hameroff who confirmed (2014b) that he and Penrose are now making an identity claim.

Such an identity claim is highly problematic for a number of reasons. For one thing, "consciousness" is a highly ambiguous term, a fact that Hameroff and Penrose readily acknowledge.

> Consciousness implies awareness: subjective, phenomenal experience of internal and external worlds. Consciousness also implies a sense of self, feelings, choice, control of voluntary behavior, memory, thought, language, and (e.g. when we close our eyes, or meditate) internally-generated images and geometric patterns. But what consciousness actually *is* remains unknown. (Hameroff and Penrose, 2014, 39)

Here I take Hameroff and Penrose to have recognized the same two uses of "consciousness" that I have distinguished, consciousness qua experience and consciousness qua experiencing subject. However, it is not clear which use is intended by the claim that consciousness *is* the orchestrated quantum state reduction given by OR. Is consciousness qua experience identical to the orchestrated quantum state reduction given by OR; or, is consciousness qua experiencing subject identical to the orchestrated quantum state reduction given by OR?

Suppose that I am experiencing an afterimage after gazing at a tomato for a sufficient length of time. Am I, this consciousness (qua experiencing subject), that which is identical to the orchestrated quantum state reduction given by OR; or, is the color of my afterimage, an aspect of my ongoing consciousness (qua experience) that which is identical to the orchestrated quantum state reduction given by OR?

Could it be both? Well, unless the experiencing subject is identical to the color of an afterimage it experiences, they could not both be identical to the same orchestrated reduction event.

Certainly, we need a more precise statement of the identity claim(s) being made by Orch OR.

#### §6.3.1.1 Identity vs Generation

Orch OR makes what I will call a generation claim. According to Hameroff and

Penrose (2014), Orch OR events *result in*: “moments of conscious awareness and/or choice” (p. 39), “a moment of consciousness” (p. 54) or “moments of conscious experience” (p. 59).

How are we to understand the generation claim in a way that is consistent with the identity claim, that consciousness ***is*** the orchestrated reduction event?

Simply combining them yields:

> **[OOR-1] Consciousness *is* the orchestrated reduction event that results in a moment of consciousness.**

How are we to understand this?

One could ask whether [OOR-1] is a statement about consciousness qua experience or consciousness qua experiencing subject; and, for that purpose, it may help to clarify the generation process.

> The Orch-OR scheme adopts DP [the Diósi–Penrose theory of objective reduction] as a physical proposal, but it goes further than this by attempting to relate this particular version of OR to the phenomenon of consciousness. Accordingly, the ‘choice’ involved in any quantum state-reduction process would be accompanied by a (miniscule) proto-element of experience, which we refer to as a moment of proto-consciousness, but we do not necessarily refer to this as actual consciousness for reasons to be described. (Hameroff and Penrose, 2014, 53)

Moments of *actual* consciousness only occur as a result of orchestration.

> If, however, a quantum superposition is (1) ‘orchestrated’, i.e. adequately organized, imbued with cognitive information, and capable of integration and computation, and (2) isolated from non-orchestrated, random environment long enough ... then Orch OR will occur and this, according to the scheme, will result in a moment of consciousness. (Hameroff and Penrose, 2014, 54)

In summarizing their proposal, Hameroff and Penrose say that each Orch OR event

> … is accompanied [by] a moment of *proto-consciousness*. These events would be thought of as the elemental constituents of ‘subjective experience’, or qualia, but the vast majority of such OR events act without being part of some coherent organized structure, so that the relevant material is normally totally dominated by random behavior in the entangled environment. Accordingly, there would normally be no significant experience associated with these ubiquitous proto-conscious events. Yet, these moments of proto-consciousness are taken to be the primitive ingredients of actual full-blown consciousness, when they are appropriately orchestrated together into a coherent whole. (Hameroff and Penrose, 2014, 70-71)

From this it seems clear that Hameroff and Penrose are talking about consciousness qua experience when they say that Orch OR results in moments of conscious experience.[66]

---

66 For the moment, I set aside objections based on claims about the orchestration process itself. For example, it is never explained how the orchestration process “imbues” a quantum superposition with cognitive significance; although, it seems likely that any explanation would involves the controversial claim Orch OR makes that, in an orchestrated reduction event, the superposition does not collapse randomly. Instead, outcome selection is “influenced by what Penrose termed ‘noncomputable Platonic values’ embedded in fundamental spacetime

Are they also talking about consciousness qua experience when stating the identity claim, consciousness is the orchestrated reduction-event? It would appear so; at least, such a reading is consistent with their claim, “Consciousness has often been argued to be a sequence of discrete moments.” (Hameroff and Penrose, 2014, 41)

That position reduces [OOR-1] to an absurdity: a moment of consciousness *is* the orchestrated reduction event that results in a moment of consciousness.

Even if Hameroff and Penrose find some way to avoid that absurdity, they face other problems. Let us set aside the difficulties that arise from the conjunction of the identity and the generation claims and consider only the identity claim. Would the possibility that consciousness qua experience is discrete rather than continuous support or undermine an identity claim?

Hameroff and Penrose write that Stroud's perceptual moment theory “described consciousness as a series of discrete events, like sequential frames of a movie” (2014, 41; Stroud, 1956). For present purposes, I'll assume that Stroud is right.[67] We (appear to) experience continuous experiencing; and, science may be telling us that this appearance is at least partially illusionary; but, the occurrence of this illusion remains a genuine experiential phenomenon, a phenomenological reality to its experiencer, something the science and philosophy of consciousness must eventually explain.

The question at hand is whether a philosopher can explain the occurrence of an illusion by claiming that there is an objectively real, physical phenomenon that is self-identical to an illusionary appearance to an experiencing subject.

In my view, that is not possible.

The experiential phenomenon, the appearance, has a different modes of existence from the physical phenomenon; so, this answer does not depend on whether the physical phenomenon in question is a neural firing pattern, a single orchestrated reduction event or a collection of reduction events appropriately orchestrated together into a coherent whole.

### §6.3.1.2 The Other Coherence Issue

If it turns out that the empirical results *require* us to accept an identity claim we will all have to live with that. If the empirical evidence does not compel the inference of identity, we are free to adopt alternate interpretations. In particular, one could tack onto the empirical results, *either* the claim that experiential phenomena are or the claim that experiential phenomena are not identical to the quantum physical phenomena with which they are correlated.

We'll just have to wait to see how that turns out.

---

geometry” (Hameroff, 2014, 132). Needless to say (but, obviously, I'm saying it anyway), this non-random collapse assumption is contrary to the assumption shared by orthodox interpretations of quantum mechanics.

67 Actually, I suspect that Stroud was on the right track. The theory of discrete experiencing certainly deserves more attention than it has received. See VanRullen and Koch (2003) for a recent defense.

Meanwhile, there is an issue of coherence concerning the identity claim that isn't easily avoided.

The identity claim – that consciousness *is* the orchestrated reduction event that results in a moment of conscious experience – seems to be in sharp contrast with Hameroff's earlier emergentive position.

> … reductionism/dualism dichotomy may potentially be resolved by views which contend that consciousness has a distinct quality, but one which emerges from brain processes which can be accounted for by natural science. (Hameroff, 1994, 91)

> Emergence implies a qualitatively new property or phenomenon which appears at a hierarchical level above the level at which rules of interaction are implemented. … Consciousness also appears to have emerged at some point in evolution. Perhaps occurring initially as a 'helpless-spectator' epiphenomenon, consciousness then assumed control of its biological environment ... The emergence of consciousness in our brains (during each conscious moment, during evolution and during the development of each human being) may be likened to new properties of materials which develop from microscopic or quantum-level events. (Hameroff, 1994, 92)

One does not naturally read these earlier passages as saying that consciousness emerges from itself – from brain processes to which it is self-identical; so, I took them to be stating a non-identity claim.

After receiving Hameroff's confirmation that Orch OR is now making an identity claim, I reviewed earlier publications for indications of such a claim. I found several, the most dramatic of which is from a paper by Hameroff and Depak Chopra:

> Penrose also suggests that each OR, or self-collapse – essentially a ripple or quantized annealing in fundamental space-time geometry – results in a moment of conscious experience.

> This is in direct contradistinction to the Copenhagen interpretation in which consciousness is outside science, externally *causing* reduction by observation. In Penrose OR, consciousness ***IS*** reduction (a particular type of reduction). Thus Penrose OR is the only worldview incorporating consciousness into the universe. (Hameroff and Chopra, 2012, 83)

The emphasis given to the 'is' in “consciousness ***IS*** reduction” makes it clear that an identity claim is being made; but, in this very paper, these authors also make other claims and speculations that seem like non-identity claims. First, of course, is the claim that each Orch OR event “results in a moment of conscious experience”. The “results in” language suggest the non-identity of cause and consequence.

Secondly, after noting that the neural correlate of ordinary states of consciousness appears to be 40 Hz gamma synchrony, Hameroff and Chopra speculate that altered states of consciousness may be associated with EEG synchrony at higher frequencies.

> At any frequency, Orch OR consciousness in the brain is occurring in fundamental space-time geometry, localized to brain neuronal microtubules and driven by metabolic processes. When the blood stops flowing, energy and oxygen depleted and microtubules inactivated or destroyed (e.g., NDE/OBE, death), it is conceivable that the

> quantum information which constitutes consciousness could shift to deeper planes and continue to exist purely in space-time geometry, outside the brain, distributed nonlocally. Movement of consciousness to deeper planes could account for NDEs/OBEs, as well as, conceivably, a soul apart from the body. (Hameroff and Chopra, 2012, 88)

Whatever the merits of such speculations, it is difficult to imagine how any of them could be consistent with the identity of consciousness and reduction events.

Whether the identity claim is a consequence of the theory or an interpretive choice made by the theorists, earlier claims should be revisited to address the possibility that they are inconsistent with the identity claim now being made.

### *§6.3.2 A Dubious Defense of Free-Will*

To be sure, Litt et al. have other objections; and, there is one that seems to me to be very strong. They write that Orch OR requires "fundamental and far-reaching revisions to quantum theory itself" (2006, 599), citing Hameroff (1998). The following is, presumably, the passage to which they object.

> … in OR (and Orch OR) the reduction outcomes are neither deterministic nor probabilistic, but 'non-computable'. The microtubule quantum superposition evolves linearly (analogous to a quantum computer) but is influenced at the instant of collapse by hidden non-local variables (quantum-mathematical logic inherent in fundamental spacetime geometry). The possible outcomes are limited, or probabilities set ('orchestrated'), by neurobiological feedback (in particular, MAPs). The precise outcome - our free-will actions - are chosen by effects of the hidden logic on the quantum system poised at the edge of objective reduction. (Hameroff, 1998, 1885)

This reliance on hidden variables seems dubious in light of recent developments in physical theory discussed above; but, I will leave the task of critiquing Orch OR from that perspective to the mathematicians who developed such interesting results from the assumption that the physicist has a free choice in the matter of which experiment to perform and how to arrange the apparatus.

The point I will address is a point that Litt et al. did not address: the revisions to quantum theory proposed by Orch OR seem to result in epiphenomenalism.

> In a volitional act possible choices may be superposed. Suppose, for example, you are selecting dinner from a menu. During preconscious processing, shrimp, sushi and pasta are superposed in a quantum computation. As threshold for objective reduction is reached, the quantum state reduces to a single classical state. A choice is made. **You'll have sushi!** (Hameroff, 1998, 1885 (emphasis supplied))

A choice is made and a moment of conscious experiencing occurs; presumably, the moment of experiencing "I'll have sushi". This sounds like the experiencing I has been reduced to being an epiphenomenon. The Orch OR event, "you'll have sushi", is made and I experience the first person translation of that fact, "I'll have sushi".

If the collapse event makes the choice, having the moment of conscious experiencing occur at the same time (a consequence of being identical to the collapse event) would seem to mean that I don't have any effect on the outcome,

telling the waitress “I'll have sushi”. The collapse event takes credit for that and all other effects; the moment of conscious experiencing seems to be there without itself having any effect at all.

In “Consciousness in the Universe”, the latest refinement of Orch OR, Hameroff and Penrose say that an Orch OR event terminating a quantum computation “would select microtubule states which could then influence and regulate axonal firings, thus controlling conscious behavior” (Hameroff and Penrose, 2014, 58).

However, the selection of a collapse outcome is a selection on the part of nature. If the collapse event occurs and a selection is made, it is not my selection. From the perspective of the experiencing I, a selection I make is a choice between two or more options open to me. I may choose between seeing movie A and seeing movie B; but, I do not (as far as I know) select the outcome of a collapse event.

It is difficult to understand how a moment of conscious experience could cause, initiate or select the outcome of a collapse event that results in that moment of conscious experience. Consequently, that moment of conscious experience seems to be an epiphenomenon.

Hameroff presents a vivid metaphor which, I am willing to stipulate, accurately depicts the operation of Orch OR events in human decision making.

> Consider a sailboat analogy for free will. A sailor sets the sail in a certain way; the direction the boat sails is determined by the action of the wind on the sail. Let's pretend the sailor is a non-conscious robot-zombie run by a quantum computer which is trained and programmed to sail. Setting and adjusting of the sail, sensing the wind and position and so forth are algorithmic and deterministic, and may be analogous to the preconscious quantum computing phase of Orch OR. The direction and intensity of the wind (seemingly capricious, or unpredictable) may be analogous to Planck-scale hidden non-local variables (e.g. 'Platonic' quantum-mathematical logic inherent in spacetime geometry). The choice, or outcome (the direction the boat sails, the point on shore that it lands upon) depends on the deterministic sail settings acted on repeatedly by the apparently unpredictable wind. Our 'free will' actions could be the net result of deterministic processes acted on by hidden quantum logic at each Orch OR event. This can explain why we generally do things in an orderly, deterministic fashion, but occasionally our actions or thoughts are surprising, even to ourselves. (Hameroff, 1998, 1885-1887)

The problem with this metaphor is that Hameroff believes that it depicts a sailor with free will; whereas, I believe it depicts a sailor whose consciousness (if any) is entirely epiphenomenal.

If my actions are the net result of deterministic neural processes being acted upon by the quantum-mathematical logic (hereafter, QML) inherent in spacetime geometry, how am I to conclude that they are freely willed *by me?* In the vivid imagery of this metaphor, *what am I?*

I am conscious; so, I am certainly not the unconscious robot-zombie (which, in the symbolism of this metaphor, is presumably my brain).

Now, I do not deny that my brain, left to its own devices, may actually *be* an unconscious robot-zombie – an instance of what U. T. Place (2000) called *the zombie within*. But, the objective of a theory of human free will is to explain how

I, *this* consciousness, have an effect on the otherwise deterministic processes of my brain (and nervous system).[68]

Could I be the QML that acts upon the otherwise deterministic processes of that robot-zombie?

This does not seem to be what Orch OR is saying. QML is one of two factors that jointly orchestrate the collapse event to which a moment of conscious experience is identical. The other factor consists of the neural phenomena that help “tune” the evolving quantum superposition. MAPs, microtubule associated proteins, "'tune' quantum oscillations, and 'orchestrate' possible collapse outcomes” (Hameroff and Penrose, 1996, 48). Together these two factors produce an outcome that is non-random but not determined by any algorithm. In the terminology of Orch OR, the outcome of (at least some) collapse events is *non-computable*.

If I am identical to a moment of conscious experience or to a sequence of such moments, I can't be identical to either of my causal ancestors – the QML and the MAP-tuning events that jointly select which collapse outcome will occur on any given occasion.

On the other hand, if I am a moment of conscious experience, I will have had no input into the 'selection' (on the part of nature) of the outcome of the collapse event of which I am a result. I did not select the collapse outcome that resulted in the occurrence of the “I'll have sushi” moment of conscious experiencing; and, similarly, I did not select the collapse outcome that resulted in the occurrence of the “I'll have shrimp” moment of conscious experiencing.

I appear to have concluded that Orch OR reduces the experiencing I to the status of an epiphenomenon subject to the illusion of agency. Of course, if Orch OR is correct, my conclusion was itself the outcome of an orchestrated objective reduction event; and, that raises further problems.

### §6.3.2.1 How Do I Store a Superposition?

Suppose I am deliberating on whether Orch OR results in an epiphenomenal experiencing I. If the shrimp-or-sushi example is a good model of the decision making process, the two possible outcomes, I take Orch OR to imply epiphenomenalism and I do not take Orch OR to imply epiphenomenalism, are represented in my brain as a quantum superposition while I am deliberating.

---

68 The word “I” is generally considered an essential indexical that (allegedly) always refers to its user (which in this case would be me, Joseph Polanik, the author of the paragraph to which this footnote is appended). Nevertheless, contemporary linguistic philosophy to the contrary notwithstanding, when I wrote “I” into that paragraph, I meant *you*, the reader. In presenting a first-person response to Orch OR, I am offering each reader the opportunity to take the “I” as a first person, self-reference. I call this reader/listener referencing use of “I” the fellow traveler or co-meditative use of “I” out of respect to Descartes who expected his readers to meditate along with him when they read the first-person prose of his *Meditations*. I do not hold his belief that everyone who thinks along with me in the first person will reach the conclusions that I reach; but, I am making a first person argument; so, if you choose to consider it at all, please let yourself experience thinking along in the first person while listening for your own response to it.

I think about this for months, going back and forth in my mind as I learn about Orch OR and consider the arguments that Penrose and Hameroff make on its behalf. What happens to the superposition of possible outcomes when I power down for the night and go to sleep? Does my brain store the superposition in memory *as a superposition*? Or does my brain simply store the two options separately along with some indication my deliberations have not yet concluded.

The latter option seems less extreme; and, avoids the need for Orch OR to explain how the superposition survives for months instead of milliseconds, However, it is not without difficulties. Suppose, after waking up the next day, I access the stored information that my deliberations remain unfinished. In resuming my deliberations, do I load stored information concerning my options into working memory?

*Do I thereby reconstruct the superposition?*

If so, it would seem that I can put my brain (or some part of it) into a superposition simply by deliberating before deciding.

At some point, this superposition self-reduces to a single outcome due to the gravitational instability of the neural representation of the alternate outcomes.

From the perspective of the experiencing I, it seems to me that I've chosen one of the two potential judgments; namely, that Orch OR reduces the experiencing I to the status of being an epiphenomenon. From the perspective of Orch Or, a moment of conscious experiencing will accompany the superposition collapse; and, presumably, I am informed as to the orchestrated outcome of my deliberations.

Assuming arguendo that Hameroff and Penrose each deny that Orch OR reduces each of them to the status of being an epiphenomenon, either I am wrong on this point or they are. The question becomes why would QML orchestrate a correct decision in some people as to whether Orch OR implies epiphenomenalism and an incorrect decision in other people?

Are people randomly selected to experience the various possible outcomes; or, is that also orchestrated; and, if so, how?

Perhaps some philosophers orchestrated into choosing to defend false positions because QML thinks that a bad philosophy is like a criminal defendant in need of a defense attorney.

Perhaps we ought to conclude that Orch OR operates to insure that all epistemological journeys are taken. That would be a bizarre validation of Cicero's indictment of philosophy: There is no idea so absurd it hasn't been advocated by some philosopher.

If it is not the case that QML operates to select all possible epistemological journeys, each occurring to at least one philosopher; then, we need some other explanation for why the brains of some philosophers are hosting Orch OR events in which faulty positions are selected.

This question about whether Orch OR implies epiphenomenalism is not the only question that may pose a problem for the Orch OR theory of decision making.

Given that we have good reasons to believe that moments of experiencing are correlated with superposition collapse events (whether orchestrated or not), one may well ask whether the correlation is due to an identity. As indicated earlier, it is my view that, until such time as the empirical research falsifies either the identity theory or the non-identity theory, we have our choice of interpretations. The identity claim and the non-identity claim would be alternate philosophical add-ons.

Suppose I am deliberating as to whether to accept the identity interpretation or the non-identity interpretation of Orch OR. After a time in which these two outcomes are superposed potentialities vying for actuality, an Orch OR event occurs and I find that I've chosen to adopt the non-identity interpretation. Penrose and Hameroff have obviously chosen the identity interpretation.

All the same questions arise as when we were dealing with the claim that Orch OR implies epiphenomenalism. How is it and why is it that superposition collapse events in the brains of some philosophers are orchestrated to result in true philosophies of consciousness in some cases and false philosophies in other cases.

Or is that random?

* * *

The Orch OR enthusiast may feel that I have not considered various arguments that Hameroff has made trying to show that Orch OR results in free will rather than epiphenomenalism; but, I wanted to draw attention to the reflexivity issue first. Orch OR purports to explain decision making; so, it must be able to explain decisions concerning Orch OR itself.

### §6.3.2.2 Retrocausality and Epiphenomenalism

Hameroff (2012) invokes retrocausality (which he calls backwards referral in time) to argue that Orch OR defends the possibility of human free will.

> Orch OR directly addresses conscious causal agency. Each reduction/conscious moment selects particular microtubule states which regulate neuronal firings, and thus control conscious behavior. Regarding consciousness occurring “too late,” quantum state reductions seem to involve temporal non-locality, able to refer quantum information both forward and backward in what we perceive as time, enabling real-time conscious causal action. Quantum brain biology and Orch OR can thus rescue free will. (Hameroff, 2012, 2)

Hameroff then mentions three lines of evidence for information referral backward in time. If anything, Hameroff probably understated the extent to which retrocausality is an acceptable concept in quantum mechanics; and, he certainly deserves credit for even mentioning the evidence for retrocausality that exists in parapsychology literature because it was too hot a topic for more mainstream journals.

All of that makes retrocausality an interesting and relevant phenomenon; but, it is not clear how retrocausality can rescue the possibility of human free will.

It may be that a reduction event selects some particular microtubule state instead of some other resulting in some behavior of which I become conscious. I will assume for the sake of argument that – in some trivial sense – behavior of which I become conscious is *conscious behavior* and causal action of which I become conscious is *conscious causal action*.

What I object to is the assumption that conscious agency only requires action or behavior of which I become conscious, which criteria is also satisfied by a epiphenomenal experiencing I. In my view conscious agency requires consciously choosing from among the options available and consciously initiating actions once a choice is made.

From this perspective, it really doesn't matter that information came to my attention after traveling backward in time. What matters is whether I make a decision based on that information. If I am simply being informed of the causal consequences resulting from a Orch OR event, then I had no part in the decision making process – irregardless of whether information traveled backwards in time to reach me.

> Each Orch OR quantum computation terminates in a moment of conscious experience, and selects a particular set of tubulin states which then trigger (or do not trigger) axonal firings, the latter exerting causal behavior. Orch OR can in principle account for conscious causal agency. (Hameroff, 2012, 14)

The reduction event *selects* ...

In a quantum physics experiment, the selection of collapse outcomes is a selection on the part of nature. It is not a selection on the part of the physicist.

If we generalize to decision making generally, the 'selection' made by the reduction event is still a selection made by nature under the influence of QML reaching up from the depths of fundamental spacetime geometry.

Orch OR is not saying either that I, consciousness, cause the collapse event or that I select the outcome of the collapse event.

Presumably the moment of conscious experience that results from the collapse event which I did not select is a moment of experiencing that occurs to me, the experiencing I. That moment of experiencing may occur as a result of the collapse event, but I accomplish nothing with it. The experiencing simply becomes aware of the outcome 'selected' by nature and QML.

In terms of Hameroff's sailboat metaphor, the robot-zombie sailor is acted upon by QML and a moment of experiencing occurs; but, the choice has already been made; so, that moment of experiencing occurs to an epiphenomenal I.

#### §6.3.2.2 Retrocausality and Superdeterminism

By itself, retrocausality – information traveling backwards through time – doesn't uniquely support the claim that humans have a free will. It can just as easily be made to support superdeterminism.

Recall what was said in the discussion of epiphenomenalism about the MIT Blackjack Team. They counted the high cards as they were dealt to learn when

the remainder of the deck had a higher than average number of high cards. They increased the size of their bets accordingly and ended up winning more money than they would have won had they bet randomly.

Retrocausality allows the universe to count cards, so to speak, for the purpose of deceiving physicists about the laws of physics.

Consider an experiment to test Bell's Theorem concerning the extent to which measurements on entangled photons are correlated. Quantum theory predicts one result; whereas, according to Bell's Theorem, hidden variable theories predict a different result – less correlation between measurements than quantum theory predicts.

Suppose that the results of an infinite series of measurements on pairs of entangled photons would be randomly distributed and that this would support hidden variable theories. Physicists don't make an infinite series of measurements. They don't measure each continuously without ever stopping. They only measure events occurring during their experiments.

The experimental design effectively assumes that the events measured during the course of the experiment are a representative sample of possible events. What if they were not? Suppose that the universe kept track of when short term anomalies in the distribution of measurement outcomes occurred and sent that information backwards in time, causing physicists to conduct their experiments only when a statistical anomaly was about to occur.

This is superdeterminism facilitated by retrocausality. Physicists would draw false conclusions about the laws of physics because the universe determined that they would conduct their experiments with non-representative data.

If Orch OR tries to defend human free will, it will need more than retrocausality. It will need something to help it avoid epiphenomenalism and superdeterminism.

### *§6.3.3 Consequences of the Penrose Solution*

According to Penrose (1989, 1994), Godel showed that, for any formal theorem-proving system of a certain easily met level of complexity, one will be able to state theorems which are not provable within that theorem-proving system but which are easily recognized as being true.

> … a specific Godel proposition neither provable nor disprovable – using the axioms and rules of the formal system under consideration – is clearly *seen*, using our insights into the meanings of the operations in question, to be a *true* proposition. (Penrose, 1989, 117)

From this, Penrose concludes that mathematical thinking goes beyond algorithmic (rule following) information processing.

> … human understanding and insight cannot be reduced to any set of computational rules. For what he appears to have shown is that no such system of rules can ever be sufficient to prove even those propositions of arithmetic whose truth is accessible, in principle, to human intuition and insight – whence human intuition and insight cannot be reduced to any set of rules. (Penrose, 1994, 65)

To fully understand Penrose's attack on computationalism, it is important to note that a quantum computer would still be algorithmic – it would simply use quantum algorithms instead of ordinary algorithms.[69]

> What can be achieved in principle by a quantum computer could also be achieved, in principle, by a suitable Turing-machine with randomizer. Thus even a quantum computer would not be able to perform the operations necessary for human conscious understanding, according to the arguments of Part 1. The hope would have to be that the subtleties of what is *really* going on when the state vector 'appears' to get reduced, rather than just the stop-gap random procedure **R** [reduction or wave function collapse], would lead us to something *genuinely* non-computable. Thus, the complete theory of the putative OR process would have to be an *essentially non-computable* scheme. (Penrose, 1994, 356)

I'm willing to assume that the brain is more like a quantum computer than a classical, digital computer; I concur with Penrose's conclusion that humans have an ability to recognize truth, an ability that exceeds formalization; and, for the sake of the argument, I concede that the experiencing of insight, of seeing or intuiting or otherwise knowing that a certain Godelian statement is true, is an instance of non-computable thinking.

What I do not concede is that this position is consistent with physicalism as that position is usually understood.

One key element of the Orch OR account of experiencing, the story of QML interacting with the deterministic processes of the brain to *orchestrate* objective reduction events, seems to make a significant concession to dualists; namely, that brain activity alone is not sufficient to account for experiencing.

Without the *orchestration* of reduction events, there would be no actual experiencing (consciousness) at all, only proto-consciousness (proto-experience); and, that orchestration is the result of the intervention of QML.

There is another echo of dualism in Orch OR. The operation of QML is reminiscent of the natural light invoked by Descartes to justify his certainty in the truth of the cogito claim. While there is a big difference between the origin of these two influences – the natural light originates with God whereas QML is inherent in fundamental spacetime geometry – there seems to be little difference in the role they play. Both intervene to help the cogitating I recognize what is true and what is false.

Arguably, however, Orch OR only defends epiphenomenalism; that's the only way I'm able to make sense of the sailboat metaphor with its robot-zombie sailor. In any case, if QML is influencing my evaluation of Orch OR, it is telling me to reject Orch OR as a disguised form of epiphenomenalism.

---

69 “... a quantum algorithm is a step-by-step procedure, where each of the steps can be performed on a quantum computer. Although all classical algorithms can also be performed on a quantum computer, the term quantum algorithm is usually used for those algorithms which seem inherently quantum, or use some essential feature of quantum computation such as quantum superposition or quantum entanglement.” (WikiPedia, 2016-07-13, Quantum_algorithm)

### *§6.3.4 The Viability of Orch OR*

Orch OR faces a number of challenges before it can claim to have explained either of the brain/consciousness relations under discussion in this thesis.

Orch OR is a speculative theory of quantum gravity; so, it must work as a theory of gravity before it can be invoked to explain how a quantum superposition in the brain's microtubule network collapses. It must make all the same predictions that general relativity makes about gravitic phenomena while still making the new predictions that explain conscious experiencing.

Orch OR has not, to my knowledge, taken into account variations in the strength of the local gravitational field affecting the brain. Was the decision making or other cognitive processes of astronauts standing on the moon impaired or enhanced by the moon's weaker gravitational field? If the strength of the local gravitational field doesn't influence the reduction of a quantum superposition in the brain, how are we to know that gravity has any effect at all on superposition collapse?

More fundamentally, it is not clear that spacetime superpositions would collapse spontaneously due to self-gravity. According to Stephen Hawkings, “the slight warping of space-time produced by the mass of a small object … will not prevent a Hamiltonian evolution with no decoherence or objective reduction” (1997, 170).

Orch OR must also show that it can work as a theory of quantum mechanics, a theory that can predict the outcome of observations on quantum systems. Standard quantum theory is a very successful scientific theory, having passed countless experimental tests. It predicts the results of a vast number of possible experiments, despite not taking the effects of gravity into account. To remain (or become) a viable attempt to explain consciousness on the basis of physics, Orch OR must make all the same predictions as standard quantum theory makes for all experiments currently known despite taking gravity into account.

To make matters worse, Orch OR rejects a significant part of orthodox quantum mechanics, the assumption that the outcome of collapse events is a random actualization of one of the potential outcomes. This element of the theory would be highly dubious even if the theory delivered what it promised, a theory of consciousness that made human free will consistent with science.

Orch OR is presented by its authors as a theory which defends free will for humans; but, this claim is undermined by the sailboat metaphor used to explain how free will is possible. Comparing humans to robot-zombie sailors affected by the QML originating in fundamental spacetime geometry is a very peculiar way of explaining how free will is possible.

Arguably, invoking QML to explain decision-making results in epiphenomenalism for the robot-zombie sailor; and, there are two other problems stemming from the introduction of this theoretical posit. First, of course, is that QML *is* another theoretical posit that the authors and advocates of Orch OR will have to justify.

The second problems stems from the claim that there is no experiencing without orchestrated reduction events. It's not just mathematical thinking or decision-

making that requires QML. Without QML there is no consciousness, only proto-consciousness; so, experiencing *any* experiential phenomenon – even something as simple as phenomenal redness – requires orchestrated reduction events which in turn require QML.

With *orchestrated* – i.e. nonrandom – reduction events so common, one naturally wonders how physicists ever got the idea that a superpositions collapses to a random selection from among the possible alternatives. Suppose a physicist is conducting a simple experiment, measuring the spin state of a particle. Standard quantum theory hold that the particle collapses randomly from a superposition of up/down to a single definite value, either spin up or spin down. But the only way the physicists can learn which outcome occurred is to be conscious of the value registered by the meter or other output device.

Becoming conscious of the outcome requires QML to non-randomly collapse the superposition in the brain of the physicist reading the output device to same value to which the particle superposition collapsed. How does QML keep track of the outcome of innumerable collapse events so it can non-randomly collapse neural superpositions to the correct value?

Finally, the identity claim that Penrose and Hameroff are now making does not sit well with the rest of the theory. Orch OR must now explain how a physical phenomenon, the quantum superposition, can be nothing other than an appearance to an experiencing subject.

## §6.4 The Dual-Aspect Reduction Event Theory, DARE

Over a period of many years, physicist Henry P. Stapp has elaborated a theory of consciousness that addresses both the brain/experience relation and the brain/subject relation. Stapp speaks about psychophysical events as reduction events having two aspects, “an experientially described aspect and also a physically described aspect” (Stapp, 2008, 14).

At one point, Stapp criticizes Jaegwon Kim for what amounts to a strawman argument, criticizing the form of dualism “advanced by Descartes during the seventeenth century instead of in the dual-aspect reduction-event form implicated by contemporary science” (Stapp, 2008, 17-18).

In light of this passage, I will treat Stapp's theory, the DARE, as a form of dualism rather than a dual-aspect theory.

* * *

Stapp develops the implications of putting the cut where von Neumann put it, at the brain/experience interface.

> von Neumann showed how to preserve the rules and precepts of quantum mechanics all the way up to the interface with “experience”, ... von Neumann’s work brings into sharp focus the central problem of interest here, which is the connection between the properties specified in the quantum mechanical description of a person’s brain and the experiential realities that populate that person’s stream of consciousness. (Stapp, 2009, 247)

This mention of a 'connection' between experience and physical events occurring

in the brain gives us a possible correlation between physical and experiential phenomena. Stapp then shows how the two kinds of phenomena are related, how they interact, without reducing one to the other.

One key idea of Stapp's approach concerns the effect of assuming a quantum physical theory instead of a classical physical theory.

> Previously the physical state was conceived to have a well defined meaning independently of any “observation”. Now the physically described state has essentially the character of a “potentia” (an “objective tendency”) for the occurrence of each one of a continuum of alternative possible “events”. (Stapp, 2009, 247)

The consequence of this shift in perspective is that an experiential phenomenon is one aspect of a psychophysical event.

> Each of these alternative possible events has both an experientially described aspect and also a physically described aspect: each possible “event” is a psychophysical happening. The experientially described aspect of an event is an element in a person’s stream of consciousness, and the physically described aspect is a *reduction* of the set of objective tendencies represented by the prior state of that person’s body-brain to the *part* of that prior state that is compatible with the increased knowledge supplied by the new element in that person’s stream of consciousness. (Stapp, 2009, 247)

By linking the reduction (collapse) event to an increase in the knowledge of the observer, Stapp has correlated a physical event and an experiential event, the two aspects of a psychophysical event.

Since neither event is reduced to being nothing more than the other event, this approach is non-reductive. Instead, it is unifying; and, the link that unifies the two events is information. In a collapse of the wave function, information is extracted from the quantum system, information as to the post-collapse state taken by the quantum system.

Now, let us advert to a principle widely believed by physicists to be true, the principle of conservation of information. (Seife, 2006, 216).

### [PCI] Information can be Neither Created nor Destroyed

If the PCI holds across the brain/experience interface, information extracted from a quantum system by a measurement can't be destroyed. Just as the law of conservation of mass/energy allows mass and energy to be *transformed*, information may be transformed from one form to another. One may then speculate that, information physically instantiated in the brain in a quantum superposition is transformed into phenomenally instantiated information in the experience of the observer, thus becoming the new knowledge gained from the experiment or observation.

If this is so, it would be the physical basis for Chalmers' postulate of double-aspect information.

> This treatment of information brings out a crucial link between the physical and the phenomenal: whenever we find an information space realized phenomenally, we find the same information space realized physically. And whenever an experience realizes an information state, the same information state is realized in the experience's physical substrate. (Chalmers, 1996, 284)

In my view that postulate is not as speculative as Chalmers made it out to be.

### *§6.4.1 The Causal Gap*

Measurements are interventions or observations performed on a quantum system; but, the Schrodinger equation does not predict which measurements will be made, when or by whom they'll be made. Someone must choose to measure something at some time by some means. This is what Stapp calls the “causal gap” in physical theory.

> According to the orthodox formulation, these interventions [measurements] are probing actions instigated by human agents who are able to 'freely' choose which one, from among various alternative possible probing actions, they will perform. (Stapp, 2007, 22)

Quantum theory is now about “consciously chosen intentional actions and their experienced feedbacks”. Whereas classical physics could avoid mentioning consciousness, “Quantum theory, on the other hand, needs something to fill a specific causal gap, and provides a means for the mind to fill it”. (Stapp, 2007, 130)

However, if consciousness is causally effective, it would have a property, the property of being able to cause a certain range of effects. A causally effective consciousness is a property bearer; hence, a substance in that most basic sense. Consequently, it is most natural to consider Stapp's theory as a form of interactive substance dualism.

I should note that, based on his comments in response to an email of mine replying to someone else on the *Journal of Consciousness Studies* email list, Stapp would likely reject this conclusion. Excepts from this exchange were included in his Mindful Universe.

> Polanik: If there is experience occurring, then to what is that experience occurring?
>
> Stapp: To what? I guess the correct question is: In what? And the answer is: in a stream of consciousness!
>
> Polanik: [Unless you choose to deny that nothingness has no properties,] it follows that there is something real to which experiences occur.
>
> Stapp: The idea that there is some physical structure 'to which experiences occur' goes far beyond what science says. Nor does science tell us that there is some immaterial entity 'to which experiences occur'. Each experience occurs in a stream of conscious[ness] which is an aspect of nature's psychophysical process. It appears 'to' a 'person' only by virtue of the fact that a person is, *actually*, according to this theory, a stream of psychophysical events, and each experience – of the kind we are considering – belongs to some such stream. The verbal statement that an experience 'occurs to' the stream to which it belongs, suggests an ontological separation that the theory does not entail or embrace. (Stapp, 2007, 132)

I'm not sure what counts as an ontological separation for Stapp; but, in my view, there is a distinction to be made between the experiencing I and the experiential phenomena it experiences. I am not identical to the afterimage I experience. I experience the afterimage; whereas, it does not experience anything at all. I persist even as it fades away.

As was noted in discussing Searle's theory of intentional actions, Stapp holds that the conscious agents that Searle thought were necessary (but which couldn't be merely a bundle of experiences) are the “... conscious agents that first choose a physical probing action, then initiate it, and finally register the response to the chosen action” (Stapp, 2010, 8).

The result of a probing action is something in the experiencer's stream of experiences which supplies the new information extracted from the quantum system. In such circumstances, it certainly sounds as if the conscious agent that chooses, initiates and registers the result of a probing action is that to which experiences occur.

In discussing Wigner's handling of the interaction between experience and brain activity, Stapp writes

> This solution is in line with Descartes' idea of two 'substances', that can interact in our brains, provided 'substance' means merely a carrier of 'essences'. The essence of the inhabitants of res cogitans is 'felt experience'. They are thoughts, ideas, and feelings: the realities that hang together to form our streams of conscious experiences. But the essence of the inhabitants of res extensa is not at all that of the sort of persisting stuff that classical physicists imagined the physical world to be made of. (Stapp, 2007, 167)

With two essence carriers, there is clearly dualism some sort of here. The crucial question is whether the experiencing I of a given stream of experiences is the initiator of the voluntary actions it appears to initiate. If so, there are two property bearers, two substances in the key philosophical sense.

Stapp explicitly acknowledges dualism (of an unspecified type).

> The quantum model of the human person is essentially dualistic, with one of the two components being described in psychological language and the other being described in physical terms. (Schwartz, Stapp and Beauregard, 2005, 16)

However, he seems to reject the interpretation of his position as substance dualism. He cites William James' dictum: “The thought itself is the thinker” in support of his view that “No new kind of entity need be doing the choosing” (2007, 133).

In my view, this defense doesn't address the crucial question: How can I make a choice without being an entity of *some* kind? If I have to be an entity of some kind to make a choice and no new kind of entity needs to be involved, which previously known kind of entity is doing the choosing? Stapp doesn't say.

In my view, nothing unreal is capable of making choices; therefore, non-entities can't make choices. If I make choices I must be something rather than nothing at all; more specifically, I must be *something capable of making choices*. If I am capable of making choices undetermined by the prior state of the universe (including my brain), I can't be identical to my brain. How would the brain alone cause (initiate) actions uncaused by its prior states?

If I am not identical to my brain, I must be … something else. What that something else is remains obscure; but, if I am causally efficacious, I have a property. Specifically, according to Stapp's theory, I would have the properties of

1. Being able to make choices undetermined by the prior state of my brain;

and,

2. Being able to initiate actions to achieve my objectives.

If I have these properties, I am a property bearer – a substance.

Consequently, the DARE is best understood as a form of interactive substance dualism different from the Cartesian form of interactive substance dualism.

## *§6.4.2 The Limited Impact of Consciousness on its Brain*

According to Stapp, the von Neumann interpretation of quantum mechanics ...

> ... takes the physical system S upon which the crucial process 1 acts to be precisely the brain of the agent, or some part of it. Thus process 1 describes here an interaction between a person's stream of consciousness, described in mentalistic terms, and an activity in their brain, described in physical terms. (Schwartz, Stapp and Beauregard, 2005, 1318)[70]

Process 1 actions are the free choices of the experimenter as to which experiments to perform and how to set up the experiment. The choices are free because they are undetermined and they seem to follow from our mental processes. It is impossible *in principle* to determine the cause of such a choice.

> [If] one does seek to determine the cause of the 'free choice' within the representation of the physical brain of the chooser, one finds that one is systematically blocked from determining the cause of the choice by the Heisenberg uncertainty principle, which asserts that the locations and velocities of, say, the calcium ions, are simultaneously unknowable to the precision needed to determine what the choice will be. (Schwartz, Stapp and Beauregard, 2005, 12)

In Stapp's model, consciousness has a limited impact on its brain

> To minimize the input of consciousness, and in order to achieve testability, we propose to allow mental effort to do nothing but control 'attention density', which is the rapidity of the process 1 events. This allows effort to have only a very limited kind of influence on brain activities, which are largely controlled by physical properties of the brain itself.
>
> …
>
> In this model all significant effects of consciousness upon brain activity arise exclusively from a well-known and well-verified strictly quantum effect known as the 'quantum Zeno effect' (QZE). (Schwartz, Stapp and Beauregard, 2005, 12)

This way of impacting the brain is enough.

> However, the study of effortfully controlled intentional action brings in two empirically accessible variables, the intention and the amount of effort. It also brings in the important physical QZE. (Schwartz, Stapp and Beauregard, 2005, 12)

---

70 Process 1 is the probing action chosen by the experimenter, Process 2 is the deterministic evolution of a quantum system according to the Schrodinger equation; and, Process 3 is the response by nature to the Process 1 intervention into a quantum system evolving according to Process 2

## §6.5 Comparing the Candidates

The candidate theories I have been considering each take wavefunction collapse to be a real physical process (somehow) linked to conscious experiencing; but, they differ as to what causes the collapse. For Orch OR it is quantum gravity, making it an objective reduction theory. For DARE and SMCT it is consciousness, making them subjective reduction theories.

They also differ as to the relation assumed between consciousness (qua experience) and collapse. Orch OR (as most recently formulated) assumes that the brain/consciousness relation is identity; whereas, DARE and SMCT assume a non-identity relation.

Each theory assumes that brain activity processes quantum information. Orch OR asserts that this information is conveyed to the brain via the microtubule network. SMCT asserts that neural spin networks are the key factor. DARE is agnostic as to the means by which the an information bearing superposition arrives at the point of collapse.

DARE is more explicit than either Orch OR or SMCT about the point of collapse being at the brain/experience interface; but, each would link the collapse event and the information extracted from it. This suggests that scientists looking to test these theories should *follow the information* to discover some way to empirically falsify one or more of these theories.

For philosophers looking to maintain consistency with the physical sciences, the suggestion is that our objective is to relate consciousness to a *quantum* brain. This point implies that SMCT, Orch OR and DARE are all aligned against those who assume that consciousness emerges at a “level” at which quantum phenomena need not be considered.

Finally, each theory supports the freedom of will against the causal closure principle; although, they differ in how this occurs. Hu and Wu say that the transcendent aspect of human consciousness is responsible for the limited free will that we have. Stapp invokes the quantum Zeno effect on behalf of the intending consciousness. Hameroff invokes the possibility of information traveling backwards in time; although, unlike DARE or SMCT, Orch OR gives every indication that it explains epiphenomenalism rather than free will.

*Do we have three theories or just two?*

As noted, both SMCT and DARE would attribute the collapse to consciousness. Consequently, as we consider the available evidence, it may help to consider SMCT and DARE as complementary theories.

SMCT is more explicit than DARE about needing two aspects of consciousness for a complete theory. Most often, only consciousness qua experience is discussed in contemporary philosophy of consciousness; but, if consciousness has the same ontological status as an afterimage, it is difficult to see how it could have the power to make choices and initiate actions independent of conditions in its brain. In essence, consciousness qua experience does not seem sufficient to account for free will.

On the other hand, while SMCT has a more elaborate theory about quantum

information processing (because specific quantum phenomena are identified), it says little about how experience arises from such information processing. DARE addresses this point by tying the information acquired via wavefunction collapse directly to experience (because the collapse occurs at the brain/experience interface).

That said, both theories would, in my view, benefit from explicitly incorporating some version of Chalmers' dual aspect theory of information.

*Which way does the evidence point?*

In §6.5.1, I'll consider the available evidence, finding very little that selectively discriminates between the candidate theories.

In §6.5.2, I'll consider a way to distinguish the DARE and SMCT from Orch OR theories by what they say concerning free will and intentional/volitional actions.

## *§6.5.1 The Experimental Evidence*

There is increasing evidence of quantum phenomena occurring in the brain; but, we don't yet have evidence that tips the balance toward one theory or the other.

Hameroff and Penrose (2014) recently presented the latest refinements to their theory and reviewed the work of an Indian team, “the Bandyopadhyay group”, that discovered a quantum phenomenon in the microtubules of the brain (Sahu et al., 2013a and 2013b).

> The resonance conductance (‘Bandyopadhyay coherence’ – ‘BC’) through tubulins and microtubules is consistent with the intra-tubulin aromatic ring pathways which can support Orch OR quantum dipoles, and in which anesthetics bind, apparently to selectively erase consciousness. Bandyopadhyay's experiments do seem to provide clear evidence for coherent microtubule quantum states at brain temperature. (Hameroff and Penrose, 2014, 55)

It is certainly encouraging that a quantum phenomenon was shown to occur in the brain's network of microtubules; and, the link between microtubular activity and anesthetic action certainly suggests that consciousness is associated with microtubular quantum phenomena. However, it has not yet been shown that this particular phenomenon performs a quantum computation.

A theory explaining avian magnetoreception via the radical-pair mechanism directly supports SMCT which makes explicit claims for the relevance of spin chemistry. One might argue that DARE is indirectly supported by such theories because the radical-pair mechanism relies on the QZE; and, that other theories invoking the QZE in a biological environment may become worthy of further consideration.

However, if the radical-pair mechanism survives as the dominant explanation of avian magnetoreception, Orch OR theorists would be challenged to explain why gravity does not interfere with the radical-pair mechanism; or, alternately, how the QZE can counteract the influence of quantum gravity.

All three theories would face challenges relating to the flow of information and to the location of the collapse event. The problem is that the radical-pair

mechanism holds that information is extracted from a quantum system in the peripheral nervous system; but, all three of our candidate theories assume that the relevant collapse events occur in the brain.

This problem is not specific to the theory of avian magnetoreception. It would occur with equal force in other cases where quantum phenomena occur in the sensory nervous system. For example, it used to be thought that the sense of smell depended on molecules in the nose that detected the shape of the incoming odorant molecules. However, recent research indicates that the odor detection system of the nose is sensitive to the vibrations of the odorant molecule. (Franco et al., 2011; Bittner et al. 2012). The prevailing explanation for olfaction is now that an electron tunneling across the receptor site measures the vibrations of the odorant molecule.[71]

If conscious experiencing involves the collapse of an information bearing superposition in the brain, our candidate theories need to explain how information extracted from a quantum system in the peripheral nervous system gets to the brain in the form of a quantum superposition.

DARE may have a fragile advantage here in that it relies on von Neumann's rejection of a preferred location for the collapse event. Stapp would be able to say that the biological system that is said to perform a measurement – the radical-pair, the electron tunneling across a receptor site or whatever – becomes entangled with the quantum system it interacts with. Other parts of the sensory nervous system that receive information from these detectors would also become entangled with them, resulting in a superposition of the detector in a fired/not-fired state.

Not until the brain/experience interface is reached would that superposition actually collapse.

This feature (if it is a feature) of DARE is a fragile advantage in that DARE would be vulnerable to evidence or arguments suggesting that the collapse actually, truly occurred at some specific point that makes it inappropriate to subject the entire body, including the brain, of the physicist to the jurisdiction of the Schrodinger equation.

For example, Fred Thaheld argues that there is an objective collapse in the eye.

> The argument is therefore made that the wave function of any superposed photon state or states is always objectively changed within the complex architecture of the eye in a continuous linear process initially for most of the superposed photons, followed by a discontinuous nonlinear collapse process later for any remaining superposed photons, thereby guaranteeing that only final, measured information is presented to the brain, mind or consciousness. (Thaheld, 2005, 113)

If information is extracted from superposed photons when a collapse event takes place in the eye, each candidate theory would need to explain how that

---

71 We might anticipate a similar revision to theories pertaining to receptor sites in the brain; for instance, the endorphin receptor sites. Opiate-like molecules can dock at the same receptor sites used by natural endorphins, presumably because their shape lets them in. But each such molecule has different effects; and, that may be due to a quantum measurement of their vibratory state.

information becomes encoded in a superposition that somehow arises as the information travels to the brain where the collapse is associated with a moment of conscious experiencing. DARE, however, would need considerable modification to incorporate a claim that von Neumann was wrong in his analysis of the measurement problem.

In any case, just as criminal investigators sometimes make progress by following the money trail, scientific investigators must follow the information.

### §6.5.1.1 Human Magnetic Sensitivity

As noted above, the radical-pair mechanism that is often used to explain avian magnetoreception depends on the photochemical, cryptochrome. Humans have a version of cryptochrome, CRY2, in their eyes; and, some have speculated that humans may be magnetosensitive.

> Somewhat surprisingly, while a human magnetosense is not widely accepted, there is accumulating evidence to suggest that such a sense -- or at least the vestiges of it -- may exist. (Close, 2012, 2082)

The extent of human magnetoreception is an open question; but, researchers have shown that CRY2 is magnetosensitive.

Normal fruit flies are magnetosensitive due to having a version of cryptochrome in their eyes; and, a strain of *Drosophila* bred to be cryptochrome deficient lost their sensitivity to the geomagnetic field. In a transgenic experiment, the human gene for CRY2 was found to restore magnetosensitivity in cryptochrome deficient *Drosophila.* (Foley et al., 2011)

However, as these researchers themselves note, just having cryptochrome in our eyes doesn't necessarily make humans magnetosensitive.

> However, we do not yet know whether this capability is translated into a downstream biological response in the human retina. Nonetheless, the transgenic findings with hCRY2, together with its anatomical location in the human retina and previous work showing field effects on the visual system, suggest that a reassessment of human magnetosensitivity may be in order. (Foley et al., 2011, 3)

Other research indicates that cryptochrome may facilitate the human response to geomagnetic storms (Close, 2012); but, it's not clear whether information is conveyed to the brain in some manner involving cryptochrome.

### §6.5.1.2 Quantum Theory of Anesthetic Action

It has been about 160 years since the discovery of general anesthetics; but, scientists are only now discovering the mechanism of anesthetic action.

It had been previously thought that general anesthetics had some sort of biochemical influence on the proteins of the brain; but, the diversity of substances that can function as a general anesthetic poses a problem. Xenon makes an excellent anesthetic; but, being an inert gas, it has no chemistry – at least not in the conditions present in the brain. However, “xenon has physics: like

many other elements and molecules, it is capable of facilitating electron transfer between conductors” (Turin et al., 2014, E3524)

Turin and his team measured an increase in electron spin in Drosophila under the influence of anesthetic gases, a signal not seen in strains of Drosophila resistant to anesthetics.

Further work will be needed to identify the specific molecule or ion that carries the signal and to determine whether the signal is the result of an increase in total spin or a change in spin polarization. If this result is corroborated, it will provide significant support for SMCT; although, ironically enough, it would also undermine the theory of anesthetic action Hu and Wu proposed in an early paper (2001): that general anesthetics reduced the availability of oxygen to oxygen utilizing sites in the brain.

Hameroff's theory of anesthetic action holds that anesthetic gases inhibit electron mobility in hydrophobic pockets within certain proteins which in turn disrupts gamma synchrony, a good candidate for being the neural correlate of consciousness (Hameroff, 2006). In an earlier paper, the relevant proteins were identified as aromatic amino acids found in the microtubules of the brain (Hameroff et al., 2002).

Litt et al. (2006) argued that Hameroff's theory of anesthetic action based on quantum phenomena performing quantum computations has been superseded by purely biochemical theories. Hameroff's reply (2006, 2007) echoes the comment made by Turin.

> ... anesthetic gas molecules are chemically inert and do not form (bio)chemical bonds with protein targets, acting solely through quantum London forces instead. Thus, to argue that biochemical explanations account for anesthesia is a non sequitur. (2007, 1040)

Recent work (Craddock et al. (2014) and Craddock et al. (2015)) indicates that the Hameroff theory of anesthetic action is still very much alive; and, the connection to Orch OR is readily apparent. Further supporting evidence was found by Emerson et al (2013) who reported that the effectiveness of one particular anesthetic, 1-aminoanthracene, was reduced by the administration of a tubulin stabilizing agent.

* * *

It's possible that further research will undermine the spin-based theory of anesthetic action favored by Hu and Wu or the electron mobility theory favored by Hameroff. It's also possible (and, in my view, more likely) that further research will narrow the difference between a theory based on electron mobility and a theory based on the spin of electrons in motion.

DARE makes no claims about anesthetic action.

### §6.5.1.3 The Weight of the Evidence

Evidence that quantum phenomena occur in the brain and nervous system is

mounting; and, one may plausibly argue that information is acquired by these quantum effects. However, much more research will be needed to establish that information reaches relevant sites in the brain in the form of a quantum superposition; and, that something interesting related to the generation of subjective experience occurs when information bearing superpositions collapse.

In the next subsection, I will suggest a line of investigation that should be feasible now or in a few years and which may be able to selectively discriminate between Orch OR on the one hand and DARE or SMCT on the other these competing theories based on what they say about intentional action.

## *§6.5.2 Wanted: The Quantum Signature of Intentional Action*

Ordinarily, I don't have to think about focusing my eyes. If I look at something, my eyes automatically assume the position necessary for the correct focal length.

Not all eye movements are volitional, of course. If I hear a sudden noise, I'll automatically turn toward the source of the sound. If I see a sudden movement in my peripheral vision, my eyes will shift; and, perhaps, my head will turn to focus on whatever moved. When this happens, I don't feel as though I have moved my eyes. I would say that the new stimulus caught my attention; and, that my nervous system moved my eyes in response.

Although the brain often moves the eyes automatically, with no input from the experiencing I, it seems intuitively obvious that I sometimes move my eyes intentionally. As I am typing this passage, it occurs to me that I should provide a simple example of intentionally moving my own eyes, an example that any reader could easily duplicate while reading this passage. While still staring at the screen as I type, I consider items that I know to be in my environment despite being outside my field of vision at the moment. I decide to look at the doorknob; and, my eyes move accordingly, assisted by a small movement of my head which I did not anticipate. I am now looking at the doorknob as I type.

To me, it seems that I chose an object on which to focus my attention; and, that my eyes moved accordingly. Now, while I know that eye movements are controlled by eye muscles, I am not aware of sending any signals to my eye muscles commanding them to move in a specific way. Nevertheless, I assume that my eyes moved in response to a signal from my brain.

*Did I move my eyes?*

For identity theorists, this is a simple question. The experiencing I is identical to its brain; so, if the brain moved its eyes, the experiencing I moved its eyes. But, I am a non-identity theorist; so, the question is more complex.

As I see it, there are two possibilities: epiphenomenalism and interactionism.

It could be that my brain not only moved my eyes without any input from me but also created the illusion that I chose what to look at and that my brain responded by moving my eyes in accordance with my intent. This view may seem palatable to physicalists who are not identity theorists because an epiphenomenal I does not challenge the causal closure principle most physicalists eagerly embrace. But, I reject the causal closure principle; so, I have no motivation to adopt

epiphenomenalism.

Alternately, it could be that my intent – to look at the doorknob – triggered a response in my brain which then signaled my eye muscles to move as required. In such circumstances, I would say that I did move my eyes. In my view, I am justified in taking credit for the actions of my brain when I initiated those actions.

Naturally, I want to know *how* my intentionality triggered, generated or otherwise caused a response from my brain. I am looking for the quantum signature of intentional action.

#### §6.5.2.1 The Experience of Viewing AutoStereograms

Stapp puts considerable emphasis on mental effort, striving for a period of time to make a decision and to achieve some objective. If DARE is correct, the QZE should be detectable during the time this effort is applied.

The conceptually simple experiment of deciding to look at a doorknob and finding that my eyes move to make it possible to see the doorknob didn't involve any noticeable effort on my part once the decision was made. On the other hand, the perception of three dimensional images in autostereograms seems to involve intentional action and mental effort over a period of time, making it a good test of the DARE theory.

In a traditional stereogram, slightly different perspectives of a single image are presented to each eye via a device known as a stereoscope. The brain extracts depth cues from the two images and combines them into a 3-D image. Single image stereograms, also known as autostereograms combine both perspectives into a singe image; thereby allowing the 3-D image to appear without the need for a stereoscope or special glasses.

In one kind of autostereogram, a random dot autostereogram, the image printed on the page appears at first glance to be a meaningless collection of spots or irregularly shaped splotches of color. Left and right perspectives are separated and depth cues are extracted when the printed image is viewed after converging or diverging the eyes (cross-eyed and wall-eyed vision respectively).

While most such autostereograms are designed to be seen with wall-eyed vision, the 3-D effect can usually be seen with cross-eyed vision as well; although, height and depth will be reversed. For example, in Baccei and Smith (2002), a collection of autostereograms, the image on page 31 will look like a raised conical mound in walleyed vision or a hole scooped out of the background in cross-eyed vision. I can see the mound or the hole as I choose – or so it seems to me.[72]

But seeing *any* 3-D image in an autostereogram takes effort; so, it's not like the bistable faces/vase image where it is hard to avoid seeing either the faces or the vase. An autostereogram is *tri-stable*. When I open a book containing such images, I initially see the flat 2-D image; and, I can focus on that for as long as I

72 In my own experience, a 3D image that is largely a simple geometric shape (e.g. a slice of pizza) is more easily reversed than a more complex image.

like, admiring the image as if it were an example of abstract art.

Presumably, when I first open a book of autostereograms, my brain adjusts the focal distance of my eyes based on the distance from my eyes to the page on which the image is printed. Once I make a choice to see either the wall-eyed or the cross-eyed 3-D image, I must *initiate* the eye movements necessary for seeing the illusionary 3-D image. To see the wall-eyed image, the focal distance of the eyes must be greater than the distance to the page. To see the cross-eyed image, it must be less.

The eyes must be moved before the image associated with a different focal length can be seen. Sometimes it is a bit of a struggle to get my eyes to focus in front of or behind the page but they will usually co-operate after a while. Once seen, the 3-D effect is stable; meaning, I don't have to further strain to keep my eyes in position to continue seeing the illusory image. Unlike with a gestalt diagram which will undergo an involuntary perceptual reversal eventually, there is no fatigue effect that I've noticed with autostereograms. One could speculate that the coherence of each image establishes a feedback loop to keep the eyes focused with the right focal distance to see the illusionary 3-D image.

Of particular interest is that, while observing the 2-D image that appears when I first look at an autostereogram in a book, I appear to be able to *choose* when to switch to seeing the 3-D view and whether to switch to seeing the cross-eyed version or the wall-eyed version of the 3-D image. Let's say I go for the wall-eyed version and succeed. I then appear to be able to choose when to switch to the cross-eyed version of the 3-D image.

This raises a number of questions of philosophical interest, among which are

1. Whether the decision to switch from 2-D viewing to wall-eyed viewing to cross-eyed viewing is my decision or whether my experience of deciding is an epiphenomenal illusion; and,
2. Whether striving to move my eyes initiates nervous system activity undetermined/uncaused by antecedent nervous system activity.

According to Stapp, the intent to act activates a *template for action* which I take to be a pattern of behavior learned through trial and error. In the case at hand, I'll assume that the template for action is the pattern of eye muscle movements necessary to move the eyes into the position required for me to see what I want to see. I would then hypothesize that the quantum state conveying the information bearing signal proceeds from an area of the brain associated with planning and executing actions, the prefrontal cortex, to an area that controls movements of the eye muscles and to the eye muscles themselves.

To see a 3-D image in an autostereogram, the eye muscles must be made to move the eyes in an unnatural way; but, as I mentioned during the discussion of epiphenomenalism, I really don't know how to move my eyes. My brain knows. All I know is that, if I intend an action that requires an eye movement, my eyes usually move in accordance with my intent.

At this point I can move my eyes around the image as if looking at various points on the background without losing the 3-D effect in the center of the image. I can

feel my eyes moving when I do this; but, this happens without any of the striving initially required to see the 3-D image.

Now, particularly with an unfamiliar autostereogram, it takes some time to move the eyes into position. To prevent a signal that is attempting to trigger the template for action from being dissipated during that time, one might invoke the quantum Zeno effect; theorizing that continued attention on the intended result prevents the quantum system conveying the signal from changing its state – preventing me from getting distracted.

Presumably, there is some feature, a default circuit of sorts, in the brain which focuses the eyes based on depth cues without conscious effort on my part. Looking at the 2-D image presented in a book of autostereograms is effortless.

Does this not indicate that experiential phenomena can have causal effects? Once I see the 3-D image, I stop striving to alter the focal distance of my eyes.

#### §6.5.2.2 The Outgoing Signal

According to Stapp, the quantum Zeno effect, QZE, is involved in the means by which my choice is conveyed to my brain. How would this work?

In the QZE, repeated observations or measurements of a quantum system tends to maintain its state. In the case of spin, for example, a particle's spin is undefined until measured because it exists in a superposition of both possible values until measured, when it randomly collapses to a single value. If, after being measured, a significant length of time passes before the particle's spin is again measured, there will be no correlation between the two measurements. But, as the length of time between the measurements decreases, the correlation between the measurements increases until, if the measurements are done rapidly enough, the particle may be said to maintain its state while under observation.

So, assuming that my decision is represented by the spin state of an ensemble of subatomic particles, maintaining my focus on my decision would count as an observation of the ensemble that instantiates that information; thereby preventing its decay into random noise.

For the sake of argument, I'll hypothesize that the signal to initiate a neuro-procedure, any neuro-procedure, is sent *from* a region related to decision making *to* the region responsible for triggering the desired neuro-procedure.

When the signal arrives at its destination, that location in the brain would be responsible for interpreting it and issuing a command to initiate a neuro-procedure. In the case we are considering, depending on my decision, the resulting command would be either “focus for wall-eyed vision” or “focus for cross-eyed vision” suitably expressed in the form of physically instantiated information. The signal would then be routed to the eye muscles themselves which would relax or contract in response to the signal. And my eyes would move accordingly.

Since Stapp's theory stresses binary choices, one might then speculate that the signals in question would each consist of an ensemble of quantum systems that could be in one of two orientations; for example, particles with two possible spin

states, spin-up or spin-down. One might further speculate that the two signals are each in some sense the reverse of the other. For example, it may not be unreasonable to hypothesize that the difference between the signal for cross-eyed viewing and the signal for wall-eyed viewing is that particles in the spin-up condition predominate in one signal and particles in the spin-down condition predominate in the other signal. But, of course, the information as to my intent may be coded in some other way.

Thus, while both signals would involve electrical activity in the prefrontal cortex (where, presumably, the choice is made between wall-eyed viewing and cross-eyed viewing) and a response in the eye muscles, the message that the electrical activity transmits may be carried by electron spin or some other quantum phenomenon that would be subject to the QZE.

The signal consisting of some quantum system subject to the QZE might proceed from the area of the brain associated with decision making directly to the eyes; but, it seems more likely that the signal would first go to the area of the brain that controls the eye muscles without conscious involvement and only then to the eyes. In either case, I'll call this the outgoing signal.

The hypothesis is that the conscious intent to focus for 3-D viewing is imposed on the system that adjusts the focal distance of the eyes and the object of vision.

Without an intentional override, that system would automatically focus the eyes based on visual cues as to the distance between the eyes and the page of the book of autostereograms that I am looking at. If this turns out to be the case, it would count as a case where a neuro-procedure was triggered by a signal indicating intentional action; consequently, it would be reasonable to allow consciousness to take credit for an action that was implemented by its brain.

*Objection! This experiment makes no attempt to detect consciousness.*

While I am proposing that physicists and neuroscientists conduct experiments along the lines discussed above, it's important to note that the experiment is about detecting the quantum signature of intentional action. It is *not* about directly detecting an immaterial consciousness thought to be responsible for intentional or freely willed actions.

It's unlikely that physicists will ever directly detect an immaterial consciousness; but, the argument is that an immaterial consciousness is necessary to account for the results of physics experiments; so, direct detection is not absolutely necessary. Something may be postulated to account for an effect that can not otherwise be explained. For example, physicists postulate that the vacuum is seething with innumerable virtual particles which emerge into existence but which last for so short a time that they can not, even in principle, be detected directly. They are postulated to exist because of the effect that they have on actually existing particles.

Similarly, if the quantum signature of intentional actions is found, one may reasonably postulate an intender of those intentional actions. If it can't be the brain, it must be … something else.

## *§7 Summation*

A long time ago, in an empire far away, the great statesman and part-time philosopher, Marcus Tullius Cicero, observed that there is no idea so absurd that it hasn't been advocated by some philosopher. Today, despite the passage of so many centuries, it seems that some philosophers are determined to illustrate the continuing validity of Cicero's indictment of philosophy.

In my view, Daniel C. Dennett is one such philosopher. His attempt to defend materialism by denying the existence of first-person phenomenology is absurd beyond belief; or, so it seems to me. So, eliminating eliminativism, I affirm the existence of first-person phenomenology.

Now, as it happens, I agree with Dennett about the consequences of affirming the existence of first-person phenomenology: it is a step across the Rubicon into dualistic territory. However, Dennett is quite vague as to the type of dualism that follows from affirming the existence of first-person phenomenology. And thus I began this epistemological journey, my effort to clarify the form of dualism that follows from tipping over that first domino.

We certainly have two vocabularies, one in which to talk about experiential phenomena occurring to an experiencing subject and one in which to talk about physical phenomena occurring in the brain of that subject; but, one might hypothesize that the referent of each of a pair of corresponding terms is one self-identical physical phenomenon.

However, on the natural assumption that first-person phenomenology consists of first-person phenomena, one may reject the hypothesis of physical/phenomenal identities by invoking the appearance/reality *distinction*.

It is not the case that an objectively detectable physical phenomenon is nothing other than an appearance to an experiencing subject; but, an experiential phenomenon such as phenomenal redness *is* nothing other than an appearance to an experiencing subject. The former exists or occurs in an experiencer independent way. The latter exists or occurs in an experiencer dependent way. Consequently, experiential phenomena are not identical to physical phenomena.

Thus, affirming the existence of first person phenomenology implies that there are two distinct kinds of phenomena; and, in my view, this state of affairs constitutes phenomenon dualism rather than phenomenon monism.

Defending the perspective I'm calling phenomenon dualism can occur on two levels. Concerning what it says about humans, it can be defended as being more plausible than the alternatives it rejects, eliminative materialism, type-Z materialism and physical/phenomenal identity theories. Concerning the meta-level of how to classify this perspective based on what it says about humans, some effort is required to defend the claim that the the position I'm calling *phenomenon dualism* is a distinct form of dualism.

Decisive in this regard is that a phenomenon dualist is making a stronger claim than predicate dualists or conceptual dualists are making: that there are two kinds of phenomena to speak about rather than two vocabularies for speaking

about one kind of phenomena.

On the other hand, a phenomenon dualist does not necessarily make any claim about the number of kinds of property required to account for having two kinds of phenomena. Both Searle and Chalmers recognize the existence of two kinds of phenomena. However, Chalmers expects that scientists will fail to explain experience without expanding their explanatory ontology to include nonphysical properties. Searle expects scientists (biologists if not physicists) to succeed in explaining experience without such an expansion.

Consequently, we are justified in putting Searle and Chalmers in different camps; and, more generally, we are justified in recognizing phenomenon dualism as a distinct form of dualism that is stronger than conceptual dualism but weaker than property dualism.

This concludes the defense of phenomenon dualism as a way of understanding the brain/*experience* relation; but, the job is not yet done. There is still the brain/*subject* relation to be considered.

Once the brain/subject relation becomes a topic of the conversation, we again encounter the three options that we encountered when considering the brain/experience relation: eliminativism, identity theory and dualism.

Not many philosophers seem willing to explicitly take a self-eliminating position, something that, in my view, would require standing and delivering “I am nothing at all” with a straight face. The position is self-refuting when stated in the first-person.

Identity theory as to the brain/subject relation is more coherent. One could assert something like “the consciousness I seem to be is identical to the brain I seem to have” without undermining one's own assertion; but, constructing an argument for that position seems highly problematic; particularly, if one must abandon the causal closure principle in favor of a free will postulate that physicists seem to have adopted.

However, denying brain/subject identity steps across the Rubicon into dualistic territory where the road quickly forks at the question of subject causation: Is the experiencing subject is epiphenomenal or causally efficacious?

Arguments for epiphenomenal dualism have been undermined by developments in quantum physics and philosophy. Physicists are increasingly making explicit use of a free will postulate; and, some philosophers have recently begun criticizing the concept of causal closure.

Consequently, I reject epiphenomenal dualism in favor of subject causationism, the conjunction of the claims that the experiencing subject exists, is not identical to its brain and is causally efficacious at least some of the time.

On one level, the defense of subject causationism consists of our reasons for adopting each of the three constituting theses and our reasons for rejecting implausible alternatives; but, there are two further considerations.

First, In weighing the evidence and the arguments for and against the three theses, it may argued that subject causationism has not been proven or that

epiphenomenal dualism has not yet been decisively ruled out. The subject causationist may simply reply “So be it. I reject epiphenomenal dualism anyway. If I am an epiphenomenon, I am an epiphenomenon who denies being an epiphenomenon.”. How does the epiphenomenal dualist explain how some epiphenomenal subjects are made to understand the truth and others are made to believe a delusion? Is there a genetic difference that predestines some thinkers for philosophical correctness (or incorrectness) about epiphenomenal dualism? I doubt anyone really wants to go down that road.

Secondly, it may be argued that subject causationism is a version of interactive substance dualism; *and*, that this is somehow a bad thing.

I do not contest the claim that subject causationism is a version of interactive substance dualism. Given certain definitions of terms such as “substance” and “property bearer”, the thesis of subject causation is plausibly viewed as a version of interactive substance dualism.

What I contest are suggestions like “and that's a bad thing because …” that may be tacked onto the claim that subject causation is a version of interactive substance dualism. Opponents of subject causationism may hold that interactive substance dualism is a bad thing for any number of reasons; but, I will make no attempt to argue against a choice based on someone's personal preferences. Thus, I won't contest claims like “and that's a bad thing because I don't like it”.

What I contest is the claim that subject causationism is a version of interactive substance dualism; and, that's a bad thing *because it is contrary to physics*.

I've argued that subject causationism is consistent with physics; so, the objection fails. One strand of this argument concerns the use being made of a free will postulate in physics. Another consists in the convergence between Searle's view that explaining a non-epiphenomenal rational agent requires incorporating quantum indeterminism and Stapp's view that quantum mechanics requires conscious agents able to make choices and initiate actions undetermined by conditions in the brain.

It may be argued that, if subject causation is consistent with physics, it can't be dualistic because whatever is physical isn't dualistic.

If physicists are required to postulate a consciousness not identical to its brain in order to explain the results of experiments, the collapse of the wavefunction or the presence of something able to make a free choice, it is hard to imagine why such a theory would not be a physical theory. But if that consciousness is not physical the same way that a stone is physical, it is equally hard to imagine why such a theory would not be a dualistic theory.

Consider the argument that Jill can not be a bachelor because Jill is female and being female is incompatible with being a bachelor. Even if this argument is valid, there is a significant disanalogy between it and an argument that being a physical theory is incompatible with being a dualistic theory. One can't inspect the meaning of the word physical and find the provision “is not dualistic”.

If we reject the analogy, one may simply conclude that the test for being a physical theory is independent of the test for being a dualistic theory. If a theory

satisfies both tests, the result is a dualistic physicalism rather than a contradiction in terms.

In my view, subject causationism is a dualistic physicalism even if it is a version of interactive substance dualism.

And, with that, the defense rests.

* * *

## ***References:***


Almog, Joseph. (2010). Dualistic Materialism. In Robert C. Koons and George Bealer (eds.) The Waning of Materialism. Oxford: Oxford University Press. 2010. pp. 349-363.

Alter, Torin. (2007). Does Representationalism Undermine the Knowledge Argument? In Torin Alter and Sven Walter (eds.), Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism, Oxford University Press.

American Chemical Society. (2009). 'Ice That Burns' may Yield Clean, Sustainable Bridge to Global Energy Future. Online at https://www.eurekalert.org/pub_releases/2009-03/acs-tb030909.php. Accessed 2016-12-28.

Anscombe, G.E.M. (1965). The Intentionality of Sensation: A Grammatical Feature. In R.J. Butler (ed.), Analytical Philosophy: Second Series, Oxford: Basil Blackwell.

Anscombe, G.E.M. (1975). The First Person. In Samuel Guttenplan (ed.), Mind and Language. (Oxford: Clarendon Press, 1975). pp. 45-65.

Baccei, Tom and Smith, Cheri. (2002). Magic Eye: Amazing 3D Illusions. Kansas City, MO: Andrews McMeel Publishing.

Bailey, Andrew. (2006). Zombies, Epiphenomenalism, and Physicalist Theories of Consciousness. *Canadian Journal of Philosophy*, 36(4):481-509.

Baker, Lynne Rudder. (1997). Why Constitution is Not Identity. *The Journal of Philosophy*. 94(12):599-621.

Baker, Lynne Rudder. (1999). Unity without Identity: A New Look at Material Constitution. *Midwest Studies in Philosophy*, 23:144-165.

Baker, Lynne Rudder. (2008). Nonreductive Materialism. In The Oxford Handbook for the Philosophy of Mind. Brian McLaughlin and Ansgar Beckermann, eds. Oxford: Oxford University Press.

Ball, Derek Nelson. (2008). Why There Are No Phenomenal Concepts, and What Physicalists Should Do About It. Ph. D. Dissertation, University of Texas at

Austin.

Ball, Derek. (2009). There Are No Phenomenal Concepts. *Mind*, 118(472):935-962.

Balog, Katalin (2012). Acquaintance and the Mind-Body Problem. In New Perspectives on Type Identity: The Mental and the Physical, Christopher Hill and Simone Gozzano (eds.), Cambridge University Press.

Bardon, Adrian. (2005). Performative Transcendental Arguments. *Philosophia*. 33(1-4):69-95.

Barnett, David. (2000). Is Water Necessarily Identical to $H_2O$? *Philosophical Studies*, 98:99–112.

Bealer, George. (1987). The Philosophical Limits of Scientific Essentialism. *Philosophical Perspectives*, 1:289-365.

Bealer, George. (1994). Mental Properties. *Journal of Philosophy*, 91(4):185-208.

Beckermann, Ansgar. (2012). Property Identity and Reductive Explanation. In Simone Gozzano and Christopher S Hill (eds.) New Perspectives on Type Identity: The Mental and the Physical. Cambridge: Cambridge University Press, 2012. pp. 66-87.

Beisecker, Dave. (2010). Zombies, Phenomenal Concepts and the Paradox of Phenomenal Judgment. *Journal of Consciousness Studies*. 17(3-4):28-46.

Benbaji, Hagit. (2008). Constitution and the Explanatory Gap. *Synthese*, 161(2):183-202.

Benbaji, Hagit. (2007). Is there a Puzzle about Water? *Philosophical Papers*, 36(2):207-218.

Bender, Michael; Sowers, Todd and Brook, Edward. (1997). Gases in Ice Cores. *Proceedings of the National Academy of Science*, USA, 94:8343–8349.

Bieri, Peter. (1992). Trying Out Epiphenomenalism. *Erkenntnis*, 36(3):283-309.

Bittner, Eric R; Madalan, Adrian; Czader, Arkadiusz and Roman, Gregg. (2012). Quantum Origins of Molecular Recognition and Olfaction in Drosophila. *The Journal of Chemical Physics*, 137: 22A551.

Block, Ned and Stalnaker, Robert. (1999). Conceptual Analysis, Dualism and the Explanatory Gap. http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/ExplanatoryGap.html retrieved 2012-09-23.

Block, Ned. (2003). The Harder Problem of Consciousness, Petrus Hispanus Lectures, *Disputatio*, 15.

Boring, Edwin G. (1930). The Physical Dimensions of Consciousness. New York: The Century Co.

Boufidis, Stavros; Karlovasitou, Anna; Balla, Aggeliki; Sitzoglou, Kostas; Vlahoyanni, Eftimia; and Baloyannis, Stavros. (2008). REM Behavior Disorder (RBD): Demographic, Clinical and Laboratory Findings in 18 Cases. *Annals of General Psychiatry*, 7(Suppl 1):S353.

Braddon-Mitchell, David and Jackson, Frank. (2007). Philosophy of Mind and Cognition, Second Edition. Oxford, UK: Blackwell Publishing.

Busemeyer, Jerome R; Pothos, Emmanuel M; Franco, Riccardo and Trueblood, Jennifer S. (2011). A Quantum Theoretical Explanation for Probability Judgment Errors. *Psychological Review*, 118(2):193–218.

Byrne, Alex and Hilbert, David R. (2003). Color Realism and Color Science. *Behavioral and Brain Sciences*. 26(1):3-64.

Cai, Jianming and Plenio, Martin B. (2013). Chemical Compass Model for Avian Magnetoreception as a Quantum Coherent Device. *Physical Review Letters*, 111, 230503.

Cai, Jianming; Guerreschi, Gian Giacomo and Briegel, Hans J. (2010). Quantum Control and Entanglement in a Chemical Compass. *Physical Review Letters*, 104, 220502.

Calef, Scott. Dualism. *Internet Encyclopedia of Philosophy*. http://www.iep.utm.edu/dualism/ retrieved 2014-07-27.

Campbell, Neil. (2003). An Inconsistency in the Knowledge Argument. *Erkenntnis*, 58: 261–266.

Cavanna, Andrea E. and Trimble, Michael R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*. 129:564-83. http://brain.oxfordjournals.org/content/129/3/564.full-text.pdf retrieved 2015-04-23.

Cavedon-Taylor, Dan. (2009). Still Epiphenomenal Qualia: Response to Muller. *Philosophia*, 37:105–107.

Chalmers, David J. (1995a). Facing up to the Problem of Consciousness. *Journal of Consciousness Studies*. 2(3):200-219. Online at http://consc.net/papers/facing.html.

Chalmers, David J. (1995b). The Conscious Mind: In Search of a Theory of Conscious Experience. Ph.D. Dissertation. University of California, Santa Cruz.

Chalmers, David J. (1996). The Conscious Mind: In Search of a Fundamental Theory. New York: Oxford University Press.

Chalmers, David J. (2003). The Content and Epistemology of Phenomenal Belief. In Q. Smith & A. Jokic (Eds.), Consciousness: New Philosophical Perspectives, (pp. 220–272). Oxford: Oxford University Press.

Chalmers, David J. (2007). Phenomenal Concepts and the Explanatory Gap. In Torin Alter and Sven Walter (eds.), Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness. New York: Oxford University Press.

Chalmers, David J. (2010). The Character of Consciousness. New York: Oxford University Press.

Chalmers, David. (2015). Consciousness and the Collapse of the Wave Function. Talk in Gottingen. https://www.youtube.com/watch?v=DIBT6E2GtjA retrieved 2016-01-30.

Chaplin, Martin. (2015). Ortho-Water and Para-Water. http://www1.lsbu.ac.uk/water/ortho_para_water.html. Update of 2015-10-19 retrieved 2016-01-01.

Churchland, Paul M. (1985). Reduction, Qualia, and the Direct Introspection of Brain States. *The Journal of Philosophy*, 82(1):8-28.

Churchland, Paul. (1989) Knowing Qualia: A Reply to Jackson. In A Neurocomputational Perspective, MIT Press. (Republished along with a 1997 Postscript in Peter Ludlow, Yujin Nagasawa and Daniel Stoljar (eds.) There's Something About Mary. Cambridge, MA: MIT Press, 2004.)

Churchland, Paul. (1997). Postscript: 1997. In Peter Ludlow, Yujin Nagasawa and Daniel Stoljar (eds.) There's Something About Mary. Cambridge, MA: MIT Press, 2004. (References paginated as in the 2004 publication.)

Clark, Austen. (1994). I am Joe's Explanatory Gap (expanded version). http://selfpace.uconn.edu/paper/PGAP.HTM retrieved 2012-03-15.

Close, James. (2012). Are Stress Responses to Geomagnetic Storms Mediated by the Cryptochrome Compass System? Proceedings of the Royal Society B, 279(1736):2081-90.

Colbeck, Roger & Renner, Renato. (2011). No extension of quantum theory can have improved predictive power. *Nat. Commun*. 2:411. Online at http://www.nature.com/ncomms/journal/v2/n8/pdf/ncomms1416.pdf retrieved 2012-02-20.

Colbeck, Roger & Renner, Renato. (2012). Is a system's wave function in one-to-one correspondence with its elements of reality? *Phys. Rev. Lett*. 108, 150402. Online at http://arxiv.org/pdf/1111.6597v2 retrieved 2012-05-26.

Colbeck, Roger & Renner, Renato. (2013a). A system's wave function is uniquely determined by its underlying physical state. http://arxiv.org/pdf/1312.7353v1 retrieved 2014-08-10.

Colbeck, Roger & Renner, Renato. (2013b). A Short Note on the Concept of Free Choice. http://arxiv.org/pdf/1302.4446v1 retrieved 2013-07-05.

Conee, Earl. (2004). Phenomenal Knowledge. In Peter Ludlow, Yujin Nagasawa and Daniel Stoljar (eds.) There's Something About Mary. Cambridge, MA: MIT Press. (originally in Australasian Journal of Philosophy, 72:2, 136-150, 1994.)

Conte, Elio; Khrennikov, Andrei Yuri; Todarello, Orlando; Federici, Antonio; Mendolicchio, Leonardo and Zbilut, Joseph P. (2009). Mental States Follow Quantum Mechanics During Perception and Cognition of Ambiguous Figures. *Open Systems & Information Dynamics*. 16(1):85-100.

Conway, John H; Kochen, Simon. (2006). The Free Will Theorem. *Found. Phys*. 36. 1441–1473. Preprint online at http://arxiv.org/pdf/quant-ph/0604079v1 retrieved 2009-07-03.

Conway, John H; Kochen, Simon. (2009). The Strong Free Will Theorem. *Notices of the AMS*. 56(2):226-232. http://www.ams.org/notices/200902/rtx090200226p.pdf retrieved 2009-12-19.

Craddock Travis JA; Friesen Douglas; Mane, Jonathan; Hameroff, Stuart and Tuszynski, Jack A. (2014). The Feasibility of Coherent Energy Transfer in Microtubules. *Journal of the Royal Society Interface*, 11:20140677.

Craddock, Travis J. A; Hameroff, Stuart R; Ayoub, Ahmed T; Klobukowski, Mariusz and Tuszynsk, Jack A. (2015). Anesthetics Act in Quantum Channels in Brain Microtubules to Prevent Consciousness. *Current Topics in Medicinal Chemistry*, 15(6):523-53.

Crane, Tim. (1995). The Mental Causation Debate. *Proceedings of the Aristotelian Society*, Supplementary Volume 69:211-236.

Crane, Tim. (2000). Dualism, Monism, Physicalism. *Mind and Society*, 1(2):73-85.

Crane, Tim. (2003). Mental Substances. In Anthony O'Hear (ed.) Minds and Persons. Cambridge University Press.

Davis, Wayne. (2014). Implicature. In *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/properties retrieved 2014-07-09.

Daviss, Bennett. (2004). Structured Water Is Changing Models. *The Scientist*, 2004-11-08.

De Brigard, Filipe. (2014). In Defense of the Self-Stultification Objection. *Journal of Consciousness Studies*, 21(5-6):120-130.

Descartes, Rene. (1984/1641). Meditations on First Philosophy. In John Cottingham, Robert Stoothoff and Dugald Murdoch (trans.), The Philosophical Writings of Descartes, Volume II, Cambridge: Cambridge University Press.

Dennett, Daniel C. (1991). Consciousness Explained. Boston: Little, Brown & Co.

Dennett, Daniel C. (2013). On a Phenomenal Confusion About Access and Consciousness. http://consciousnessonline.com/2013/02/15/on-a-phenomenal-confusion-about-access-and-consciousness/

Diaz-Leon, E. (2008). Defending the Phenomenal Concept Strategy. *Australasian Journal of Philosophy*, 86(4):597-610.

Dorato, Mauro and Laudisa, Federico. (2014). Realism and instrumentalism about the wave function. How should we choose? forthcoming in Gao Shan (ed.) Protective Measurements and Quantum Reality: Toward a New Understanding of Quantum Mechanics. Cambridge University Press. http://arxiv.org/pdf/1401.4861v1 retrieved 2015-03-03.

Dresler M; Wehrle R; Spoormaker VI; Koch SP; Holsboer F; Steiger A; Obrig H; Sämann PG; Czisch M. (2012). Neural correlates of dream lucidity obtained from contrasting lucid versus non-lucid REM sleep: a combined EEG/fMRI case study. *SLEEP,* 35(7):1017–1020. Online at http://www.journalsleep.org/ViewAbstract.aspx?pid=28569 retrieved 2015-04-21.

Dugarte, Edgar Jose Candales. (2014). Analysis of Free Will and Determinism in Physics. http://arxiv.org/pdf/1407.1804v1 retrieved 2015-02-24.

Emerson, Daniel J; Brian P; Psonis, John; Liao, Zhengzheng; Taratula, Olena; Fiamengo, Ashley; Wang, Xiaozhao; Sugasawa, Keizo; Smith III, Amos B; Eckenhoff Roderic G. and Dmochowski, Ivan J. (2013). Direct Modulation of

Microtubule Stability Contributes to Anthracene General Anesthesia. *Journal of the American Chemical Society*, 10;135(14):5389-98.

Evnine, Simon. (2011). Constitution and Composition: Three Approaches to their Relation. *Protosociology*, 27:212-235.

Fisher, Matthew P.A. (2015). Quantum Cognition: The Possibility of Processing with Nuclear Spins in the Brain. *Annals of Physics*. 362:593–602.

Franco, Maria Isabel; Turin, Luca; Mershin, Andreas and Skoulakis, Efthimios M. C. (2011). Molecular Vibration-Sensing Component in Drosophila Melanogaster Olfaction. *Proceedings of the National Academy of Sciences, USA*, 108(9):3797-3802.

Feigl, Herbert. (1967). The "Mental" and the "Physical": The Essay and a Postscript. *Minnesota Studies in the Philosophy of Science: Concepts, Theories, and the Mind-Body Problem*, volume 2. Online at http://ditext.com/feigl/mp/mp.html retrieved 2016-03-20.

Feser, Edward. (2004). Why Searle *Is* a Property Dualist. Paper presented at the American Philosophical Association Pacific Division meeting in Pasadena, CA, March 24-28, 2004. Online at http://www.edwardfeser.com/unpublishedpapers/searle.html.

Fisher, Matthew P.A. (2015). Quantum Cognition: The Possibility of Processing with Nuclear Spins in the Brain. *Annals of Physics*, 362:593–602.

Francescotti, Robert. (2003). Statues and Their Constituents: Whether Constitution is Identity. *Metaphysica*, 4(2):59-78.

Frege, Gottlob. (1956). The Thought: A Logical Inquiry. *Mind, New Series*. 65(259):289-311. http://www.jstor.org/stable/2251513.

Foley, Lauren. E.; Gegear, Robert J. and Reppert, Steven M. (2011). Human Cryptochrome Exhibits Light-Dependent Magnetosensitivity. *Nature Communications*, 2:356.

Foster, John. (2012). Subjects of Mentality. In After Physicalism (ed.) by Benedikt Paul Gocke. Notre Dame, Indiana: University of Notre Dame Press.

Gadenne, Volker. (2006). In Defense of Qualia Epiphenomenalism. *Journal of Consciousness Studies*, 13(1-2):101-114.

Garrett, Brian Jonathan. (2000). Defending Non-Epiphenomenal Event Dualism. *The Southern Journal of Philosophy*, 38:393-412.

Gearhart, Livingston and Persinger, Michael A. (1986). Geophysical Variables and Behavior: XXXIII. Onsets of Historical and Contemporary Poltergeist Episodes Occurred with Sudden Increases in Geomagnetic Activity. *Perceptual and Motor Skills*, 62:463-466.

Ghirardi, Giancarlo and Romano, Raffaele. (2015). Is a Description Deeper than the Quantum One Possible? http://arxiv.org/pdf/1501.04127v1. Retrieved 2015-03-01.

Gibbard, Allan. (1975). Contingent Identity. *Journal of Philosophical Logic*, 4(2):187-222.

Gmeindl, Leon; Chiu, Yu-Chin; Esterman, Michael S; Greenberg, Adam S; Courtney, Susan M. and Yantis, Steven. (2016). Tracking the Will to Attend: Cortical Activity Indexes Self-Generated, Voluntary Shifts of Attention. *Attention, Perception, & Psychophysics*, 78(7):2176–2184.

Gocke, Benedikt Paul. (2012). Introduction in After Physicalism (ed.) by Benedikt Paul Gocke. Notre Dame, Indiana: University of Notre Dame Press. http://www.academia.edu/attachments/30399960/download_file retrieved 2015-01-23

Hagan, S; Hameroff, SR. and Tuszyński JA. (2002). Quantum Computation in Brain Microtubules: Decoherence and Biological Feasibility. *Physical Review E*, 65:061901.

Hameroff, Stuart R. (1994). Quantum Coherence in Microtubules: A Neural Basis for Emergent Consciousness? *Journal of Consciousness Studies*, 1(1):91-118.

Hameroff, Stuart. (1998). Quantum Computation in Brain Microtubules? The Penrose-Hameroff “Orch OR” Model of Consciousness. *Philosophical Transactions of the Royal Society of London A*, 356:1869–1896.

Hameroff, Stuart R. (2006). The Entwined Mysteries of Anesthesia and Consciousness. *Anesthesiology*, 105(2):400-12.

Hameroff, Stuart R. (2007). The Brain Is Both Neurocomputer and Quantum Computer. *Cognitive Science*, 31:1035–1045.

Hameroff, Stuart. (2012). How Quantum Brain Biology can Rescue Conscious Free Will. *Frontiers in Integrative Neuroscience*. 6:93. http://journal.frontiersin.org/Journal/10.3389/fnint.2012.00093/pdf Retrieved 2012-12-22.

Hameroff, Stuart. (2014a). Consciousness, Microtubules and "Orch-OR": A 'Space-time' Odyssey. *Journal of Consciousness Studies*, 21(3-4):126-153.

Hameroff, Stuart. (2014b). Personal Communication. 2014-08-19.

Hameroff, Stuart and Chopra, Deepak. (2012). The “Quantum Soul”: A Scientific Hypothesis. In Alexander Moreira-Almeida and Franklin Santana Santos (eds.) Exploring Frontiers of the Mind-Brain Relationship. New York: Springer. pp. 79-93.

Hameroff, Stuart and Penrose, Roger. (1996). Conscious Events as Orchestrated Space-Time Selections. *Journal of Consciousness Studies*. 3(1):36-53.

Hameroff, Stuart and Penrose, Roger. (2014). Consciousness in the Universe: A Review of the 'Orch OR' Theory. *Physics of Life Reviews*, 11:39-78.

Hameroff, S; Nip, A; Porter, M and Tuszynski, J. (2002). Conduction Pathways in Microtubules, Biological Quantum Computation, and Consciousness. *Biosystems*, 64(1-3):149-68.

Harman, Gilbert. 1990. The Intrinsic Quality of Experience, *Philosophical Perspectives*, 4. http://ruccs.rutgers.edu/faculty/pylyshyn/Consciousness_2014/Harman_IntrinsicQualityOfExperience.pdf retrieved 2014-12-20.

Hartmann, L; Düer, W. and Briegel, H.-J. (2006). Steady-State Entanglement in Open and Noisy Quantum Systems. *Physical Review A*, 74:052304.

Hawking, Stephen. (1997). The Objections of an Unashamed Reductionist. In M. Longair (Ed.), The Large, the Small and the Human Mind. Cambridge, England: Cambridge University Press. pp. 169–172.

Henderson, Richard W. (2014). Breaking the Spell: Materialism and the Qualia Intuition. *Journal of Consciousness Studies*. 21(7-8):184-192.

Herbut, Fedor. (2014). Fleeting Critical Review of the Recent Ontic Breakthrough in Quantum Mechanics. http://arxiv.org/pdf/1409.6290v1.pdf retrieved 2014-10-26.

Hill, Christopher S. (2012). Locating Qualia: Do They Reside in the Brain or in the Body and the World? In Simone Gozzano and Christopher S Hill (eds.) New Perspectives on Type Identity: The Mental and the Physical. Cambridge: Cambridge University Press, 2012.

Hintikka, Jaako. (1965). Cogito, Ergo Sum: Inference or Performance? In Meta-Meditations: Studies in Descartes by Alexander Sesonske and Noel Fleming (eds.), Belmont, CA: Wadsworth Publishing Co. pp. 50-76.

Horgan, Terence. (1984a). Functionalism, Qualia, and the Inverted Spectrum. *Philosophy and Phenomenological Research*, 44(4):453-469.

Horgan, Terence. (1984b). "Jackson on Physical Information and Qualia", *Philosophical Quarterly*, 34(135):147-152. (Republished in Peter Ludlow, Yujin Nagasawa and Daniel Stoljar (eds.) There's Something About Mary. Cambridge, MA: MIT Press, 2004.)

Howell, Robert J. (2009). The Ontology of Subjective Physicalism. *Nous*, 43(2):315–345. Preprint online at https://www.academia.edu/3392082/The_Ontology_of_Subjective_Physicalism retrieved 2015-08-15.

Howell, Robert J. (2013). Consciousness and the Limits of Objectivity: The Case for Subjective Physicalism. Oxford: Oxford University Press.

Hu, Huping and Wu, Maoxin. (2001). Mechanism of Anesthetic Action: Oxygen Pathway Perturbation Hypothesis. *Medical Hypotheses*, 57(5): 619-627.

Hu, Huping and Wu, Maoxin. (2004a). Spin-Mediated Consciousness Theory: Possible Roles of Neural Membrane Nuclear Spin Ensembles and Paramagnetic Oxygen. *Medical Hypotheses*, 63(4):633-646.

Hu, Huping and Wu, Maoxin. (2004b). Action Potential Modulation of Neural Spin Networks Suggests Possible Role of Spin. *NeuroQuantology*, 2(4):309-317.

Hu, Huping and Wu, Maoxin. (2006a). Nonlocal Effects of Chemical Substances on the Brain Produced through Quantum Entanglement. *Progress in Physics*, 3:20-26.

Hu, Huping and Wu, Maoxin. (2006b). Evidence of Non-Local Physical, Chemical and Biological Effects Supports Quantum Brain. *NeuroQuantology*, 4(4):291-306.

Hu, Huping and Wu, Maoxin. (2007). Evidence of Non-Local Chemical, Thermal

and Gravitational Effects. *Progress in Physics*, 2:17-24.

Hu, Huping and Wu, Maoxin. (2008). Concerning Spin as Mind-Pixel How Mind Interacts with the Brain Through Electric Spin Effects. *NeuroQuantology*, 6(1):26-31.

Hu, Huping and Wu, Maoxin. (2010). Consciousness-Mediated Spin Theory: The Transcendental Ground of Quantum Reality. *Journal of Consciousness Exploration and Research*. 1(8):937-970.

Hu, Huping and Wu, Maoxin. (2013). Human Consciousness as Limited Version of Universal Consciousness. *Journal of Consciousness Exploration and Research*, 4(1):52-68.

IUPAC (1997). Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by A. D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford. XML on-line corrected version: http://goldbook.iupac.org (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins.

Jackson, Frank. (1977). Perception. Cambridge: Cambridge University Press.

Jackson, Frank. (1982). Epiphenomenal Qualia, *Philosophical Quarterly*, 32:127-36. (Republished in Peter Ludlow, Yujin Nagasawa and Daniel Stoljar (eds.) There's Something About Mary. Cambridge, MA: MIT Press, 2004.)

Jackson, Frank. (1986). What Mary Didn't Know, *The Journal of Philosophy*, 83(5):291-295.

Jackson, Frank. (1998/2004). Postscript on Qualia. In Peter Ludlow, Yujin Nagasawa and Daniel Stoljar (eds.) There's Something About Mary. Cambridge, MA: MIT Press, 2004, pp. 417-420. (Originally in Frank Jackson, Mind, Method and Conditionals. London: Routledge.)

Jackson, Frank. (2003). Mind and Illusion. *Royal Institute of Philosophy Supplement*, 53:251-271. (Also in Peter Ludlow, Yujin Nagasawa and Daniel Stoljar (eds.) There's Something About Mary. Cambridge, MA: MIT Press. 2004.)

Jackson, Frank. (2007). "The Knowledge Argument, Diaphanousness, Representationalism" in Torin Alter and Sven Walter (eds.), Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism, Oxford University Press.

Jackson, Frank. (2012). In Defence of the Identity Theory Mark I. In Hill Christopher & Gozzano Simone (eds.), New Perspectives on Type Identity: The Mental and the Physical. Cambridge University Press.

James, William. (1983/1890). The Principles of Psychology. Cambridge, MA: Harvard University Press.

Jammer, Max. 1974. The Philosophy of Quantum Mechanics: The Interpretations of Quantum Mechanics in Historical Perspective. New York: John Wiley & Sons.

Janssens, Bas and Maassen, Hans. (2006). Information Transfer Implies State Collapse. *J. Phys. A: Math. Gen*. 39:9845-9860. Online at arXiv:quant-ph/0602140v1 retrieved 2008-12-29.

Johnston, Mark. (1992a). Constitution Is Not Identity. *Mind, New Series*.

101(401):89-105.

Johnston, Mark. (1992b). How to Speak of the Colors. *Philosophical Studies*, 68:221-263.

Johnston, Mark. (1997a). Manifest Kinds. *The Journal of Philosophy*, 94(11):564-583.

Johnston, Mark. (1997b). It Necessarily Ain't So. Published online, http://www.nyu.edu/gsas/dept/philo/courses/consciousness97/papers/johnston/aintso.html. Retrieved 2015-01-03.

Kim, Jaegwon. (1989). The Myth of Nonreductive Materialism. *Proceedings and Addresses of the American Philosophical Association*. 63(3):31-47. http://www.jstor.org/stable/3130081.

Kim, Jaegwon. (1976). Events as Property Exemplifications. In M. Brand & D. Walton (eds.), Action Theory. D. Reidel. 310-326. (Also in Supervenience and Mind by Jaegwon Kim (ed). Cambridge: Cambridge University Press: 33-52.)

Kim, Jaegwon. (1993). Events as Property Exemplifications. In Supervenience and Mind by Jaegwon Kim (ed). Cambridge: Cambridge University Press: 33-52.

Kim, Jaegwon. (1999). Making Sense of Emergence. *Philosophical Studies*, 95:3–36.

Kim, Jaegwon. (2005). Physicalism, Or Something Near Enough. Princeton, NJ:Princeton University Press.

Kim, Jaegwon. (2015). Personal Communication. Email of 2015-01-23.

Kobes, Bernard W. (2007). The Philosopher's Projective Error. *Philosophical Studies*, 132(3):581-593.

Koch, K. and Hepp. (2006). Quantum Mechanics in the Brain. *Nature*. 440. p. 611.

Kominis, Iannis K. (2008). Quantum Zeno Effect Underpinning the Radical-Ion-Pair Mechanism of Avian Magnetoreception. http://arxiv.org/pdf/0804.2646v1.pdf. Retrieved 2016-08-14.

Kozuch, Benjamin P. and Kriegel, Uriah. (2015). Correlation, Causation, Constitution: On the Interplay between the Science and Philosophy of Consciousness. In Steven M. Miller (ed.), The Constitution of Phenomenal Consciousness: Toward a Science and Theory. Amsterdam: John Benjamins Publishing Company. pp. 400-417.

Kriegel, Uriah. (2007a). Philosophical Theories of Consciousness: Contemporary Western Perspectives in P. D. Zelaso, M. Moscovitch and E. Thompson (Eds.) The Cambridge Handbook of Consciousness. Cambridge: Cambridge University Press. 2007. http://books.google.com/books?id=o9ZRc6-FDg8C.

Kriegel, Uriah. (2007b). Gray Matters: Functionalism, Intentionalism and the Search for NCC in Jeffrey Gray's Work. *Journal of Consciousness Studies*. 14(4):96-116.

Kriegel, Uriah. (2009). Subjective Consciousness: A Self-Representational

Theory. New York: Oxford University Press.

Lemmon, E. J. (1965). Beginning Logic. London: Thomas Nelson and Sons.

Levine, Joseph. (1993), On leaving out what it's like. In: M. Davies and G. W. Humphreys (eds.) Consciousness. Oxford: Blackwell, 121-136.

Lewis, Clarence Irving. (1929/1956). Mind and the World Order. New York: Dover.

Lewis, David K. (1966) An Argument for the Identity Theory, *Journal of Philosophy*, 63:17–25. http://www.andrewmbailey.com/dkl/Identity_Theory.pdf retrieved 2014-06-26.

Lewis, David K., (1988). What Experience Teaches. In Peter Ludlow, Yujin Nagasawa and Daniel Stoljar (eds.) There's Something About Mary. Cambridge, MA: MIT Press, 2004. pp. 77-103. (Originally in *Proceedings of the Russellian Society*, 13:29-57, 1988. Quotes paginated as in the collection.)

Lewis, Peter J. (2006). Conspiracy Theories of Quantum Mechanics. *British Journal for the Philosophy of Science*. 57(2):359–381.

Li, J. and Paraoanu, G.S. (2009). Generation and Propagation of Entanglement in Driven Coupled-Qubit Systems. *New Journal of Physics*, 11:113020.

Litt, Abninder; Eliasmith, Chris; Kroon, Frederick W; Weinstein, Steven and Thagard, Paul. (2006). Is the Brain a Quantum Computer? *Cognitive Science*, 30:593–603.

Loar, Brian, (1997). Phenomenal States (Revised). In Ned Block, Owen Flanagan and Guven Guzeldere (eds.). The Nature of Consciousness: Philosophical Debates. Cambridge, MA: MIT Press.

London, F., and Bauer, E. (1939). The Theory of Observation in Quantum Mechanics in J.A. Wheeler and W.H. Zurek (ed.), Quantum Theory and Measurement. Princeton: Princeton University Press, 1983, pp. 217-259.

Lowe, E. J. (2006). Non-Cartesian Substance Dualism and the Problem of Mental Causation. *Erkenntnis*, 65:5–23.

Lowe, E. J. (2012). Non-Cartesian Substance Dualism. In Benedikt Paul Gocke (ed.) After Physicalism. Notre Dame, Indiana: University of Notre Dame Press.

Lycan, William G. (1987). Consciousness. Cambridge, MA: MIT Press.

Lycan, William G. (2015). Representational Theories of Consciousness. In *Stanford Encyclopedia of Philosophy*. Online at http://plato.stanford.edu/entries/consciousness-representational/ retrieved 2015-06-24.

Malt, Barbara C. (1994). Water is not $H_2O$. *Cognitive Psychology*, 27(1):41-70.

McGinn, Colin. (1989). Can We Solve the Mind-Body Problem? *Mind, New Series*, 98(391):349-366.

Moore, Dwayne. (2012a). On Robinson's Response to the Self-Stultifying Objection. *Review of Philosophy and Psychology*, 3(4):627-641.

Moore, Dwayne. (2012b). Physical-Effect Epiphenomenalism and Common Underlying Causes. *Dialogue*, 51(3):397-418.

Muller, Hans. (2008). Why Qualia Are Not Epiphenomenal. *Ratio* (new series) 21(1):85-90.

Nagasawa, Yujin. (2010). The Knowledge Argument and Epiphenomenalism. *Erkenntnis*, 72:37–56.

Nagel, Thomas. (1974). What is it Like to Be a Bat? *The Philosophical Review*, 83(4):435-50.

Nagel, Thomas. (1998). Conceiving the Impossible and the Mind-Body Problem. *Philosophy*. 73(285):337-352.

Nemirow, Laurence. (2007). So This Is What It's Like: A Defense of the Ability Hypothesis. In Torin Alter and Sven Walter (eds.), Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism, Oxford University Press.

NESTA, National Earth Science Teachers Association. (2010). Temperature of Ocean Water. Online at http://www.windows2universe.org/earth/Water/temp.html. Accessed 2017-01-07.

Nida-Rümelin, Martine. (2006). Dualist Emergentism. In Brian McLaughlin & Jonathan Cohen (eds.), Contemporary Debates in Philosophy of Mind, Blackwell. Online at http://www.exre.org/assets/files/nida/Dualist%20Emergentism%20letzte%20Fassung retrieved 2011-10-15.

Nida-Rümelin, Martine. (2010). An Argument from Transtemporal Identity for Subject-Body Dualism. In Robert C. Koons and George Bealer (eds.) The Waning of Materialism. Oxford: Oxford University Press.

Nordby, Knut. (2007). What is this Thing Called Color: Can a Totally Color-Blind Person Know about Color? In Torin Alter and Sven Walter (eds.), Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism, Oxford University Press. pp. 77-83.

O'Reilly, Edward J. and Olaya-Castro, Alexandra. (2014). Non-Classicality of the Molecular Vibrations Assisting Exciton Energy Transfer at Room Temperature. *Nature Communications*, 5:3012.

Ouellette, Jennifer. (2016). A New Spin on the Quantum Brain. *Quanta*, 2016-11-02. Online at https://www.quantamagazine.org/20161102-quantum-neuroscience/.

Palmquist, Stephen. (1987). A Priori Knowledge in Perspective: (II) Naming, Necessity and the Analytic A Posteriori. *Review of Metaphysics*, 41:255-282.

Papineau, David. (2001). The Rise of Physicalism. In Carl Gillett and Barry Loewer (eds.) Physicalism and its Discontents. Cambridge, UK: Cambridge University Press. 3–36. https://www.scribd.com/doc/227799960/Papineau-D-Rise-of-Physicalism retrieved 2015-01-24.

Papineau, David. (2002). Thinking about Consciousness. New York: Oxford University Press.

Papineau, David. (2007). Phenomenal and Perceptual Concepts. In Torin Alter and Sven Walter (eds.), Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism, Oxford University Press.

Papineau, David. (2008). Explanatory Gaps and Dualist Intuitions. In Lawrence Weiskrantz and Martin Davies (eds.), Frontiers of Consciousness: The Chichele Lectures. New York: Oxford University Press.

Papineau, David. (2009). The Causal Closure of the Physical and Naturalism. In B. McLaughlin, A. Beckermann and S. Walter (eds.), Oxford Handbook of the Philosophy of Mind. New York: Oxford University Press, pp. 53-65.

Papineau, David. (2015a). Consciousness, Emergence & Mental Causation. https://www.youtube.com/watch?v=VbqkatmW-qU retrieved 2016-01-31.

Papineau, David. (2015b). Physicalism without Causal Closure. Conference presentation at Towards a Science of Consciousness, Helsinki.

Pauen, Michael; Staudacher, Alexander and Walter, Sven. (2006). Epiphenomenalism - Dead or a Way Out. *Journal of Consciousness Studies*. 13(1–2):7–19.

Penrose, Roger. (1989). The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics. Cambridge, UK: Oxford University Press.

Penrose, Roger. (1994). Shadows of the Mind: A Search for the Missing Science of Consciousness. Oxford: Oxford University Press.

Pereboom, Derk. (2015). The Material Constitution of Phenomenal Consciousness. In Steven M. Miller (ed.), The Constitution of Phenomenal Consciousness: Toward a Science and Theory. Amsterdam: John Benjamins Publishing Company. pp. 418-432.

Persinger, M.A. (1976). Transient Geophysical Bases for Ostensible UFO-Related Phenomena and Associated Verbal Behavior. *Perceptual and Motor Skills*, 43, 215-221.

Persinger, Michael A. (1979). ELF Field Mediation in Spontaneous PSI Events: Direct Information Transfer or Conditioned Elicitation? *Psychoenergetic Systems*, 3:155-169.

Persinger, M.A. (1981). Geophysical Variables and Behavior: III. Prediction of UFO Reports by Geomagnetic and Seismic Activity. *Perceptual and Motor Skills*, 53, 115-122.

Persinger, M.A. (1982). Geophysical Variables and Behavior: IV. UFO Reports and Fortean Phenomena: Temporal Correlations in the Central USA. *Perceptual and Motor Skills*, 54, 299-302.

Persinger, M.A. (1983). Geophysical Variables and Behavior: VII. Prediction of Recent European UFO Reports by Nineteenth-Century Luminosity and Solar-Seismic Measures. *Perceptual and Motor Skills*, 56, 91-95.

Persinger, Michael A. (1985). Geophysical Variables and Behavior: XXVI. A Response to Rutkowski's Critique of the Tectonic Strain Hypothesis for UFO Phenomena. *Perceptual and Motor Skills*, 60:575-582.

Persinger, Michael A. (1988). Increased Geomagnetic Activity and the Occurrence of Bereavement Hallucinations: Evidence for Melatonin-Mediated Microseizuring in the Temporal Lobe? *Neuroscience Letters*, 88:271-74.

Persinger, Michael A. and Krippner, Stanley. (1989). Dream ESP Experiments and Geomagnetic Activity. *Journal of the American Society for Psychical Research*, 83:101-116.

Persinger, Michael A; Saroka, Kevin S; Koren, Stanley A. and St-Pierre, Linda S. (2010). The Electromagnetic Induction of Mystical and Altered States within the Laboratory. *Journal of Consciousness Exploration & Research*, 1(7):808-830.

Persinger, Michael A.; Dotta, Blake. T.; Saroka, Kevin S. and Scott, Mandy A. (2013). Congruence of Energies for Cerebral Photon Emissions, Quantitative EEG Activities and ~5 nT Changes in the Proximal Geomagnetic Field Support Spin-based Hypothesis of Consciousness. *Journal of Consciousness Exploration and Research*, 4(1):1-24.

Pickel, Bryan. (2010). There is no 'Is' of Constitution. *Philosophical Studies*, 147:193–211.

Place, Ullin T. (1956). Is Consciousness a Brain Process? In George Graham, Elizabeth R. Valentine (eds.). Identifying the Mind: Selected Papers of U. T. Place. New York: Oxford University Press. 2004. pp. 45-52. (Originally in *British Journal of Psychology*, 47(1):44-56. Quotes paginated as in the collection).

Place, Ullin T. (1999). Token- versus Type-Identity Physicalism. In George Graham and Elizabeth R. Valentine (eds.), Identifying the Mind: Selected Papers of U. T. Place. New York: Oxford University Press (2004). (originally in *Anthropology and Philosophy*, 3:21–31, 1999. Quotes paginated as in the collection)

Place, U. T. (2000). Consciousness and the "Zombie-Within" – A Functional Analysis of the Blindsight Evidence. In George Graham, Elizabeth R. Valentine (eds.). Identifying the Mind: Selected Papers of U. T. Place. New York: Oxford University Press. 2004, pp. 113-137. (Originally in Beyond Dissociation, edited by Yves Rossetti and Antti Revonsuo.)

Place, Ullin T and Schneider, Steven. (2013). Identity Theories. In *A Field Guide to the Philosophy of Mind*. http://host.uniroma3.it/progetti/kant/field/mbit.htm retrieved 2013-08-31 (Date of first publication is unknown to me; citing date of retrieval instead. The page itself lists only Place as the author; but, the index to the Field Guide lists Schneider as well.)

Pothos, Emmanuel M. and Busemeyer, Jerome R. (2009). A Quantum Probability Explanation for Violations of 'Rational' Decision Theory. *Proceedings of the Royal Society B*. 276:2171–2178.

Pothos Emmanuel M. and Busemeyer Jerome R. (2013). Can Quantum Probability Provide a New Direction for Cognitive Modeling? *Behavioral and Brain Sciences*, 36(3):255-274.

Pusey, Matthew F; Barrett, Jonathan; Rudolph, Terry. (2012). On the reality of the quantum state. *Nature Physics*. 8:476. Preprint http://arxiv.org/pdf/1111.3328v2

retrieved 2012-05-26.

Putnam, Hilary. (1967). Psychological Predicates. In W. H. Capitan and D. D. Merrill (eds.), Art, Mind, and Religion. Pittsburgh: University of Pittsburgh Press.

Ritz, Thorsten; Adem, Salih and Schulten Klaus. (2000). A Model for Photoreceptor-Based Magnetoreception in Birds. *Biophysical Journal*, 78(2):707-18.

Robinson, Howard. (2008). Why Frank Should Not Have Jilted Mary. In Edmond Wright (ed.), The Case for Qualia. Cambridge, MA: MIT Press. 223-245.

Robinson, Howard. (2011). Dualism. *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/dualism/ retrieved 2014-06-26.

Robinson, Howard. (2012). Qualia, Qualities and Our Conception of the Physical World. In Benedikt Paul Gocke (ed.) After Physicalism. Notre Dame, Indiana: University of Notre Dame Press.

Robinson, William S. (1982). Causation, Sensations and Knowledge, *Mind*, New Series, 91(364):524-540.

Robinson, William S. (2004). *Understanding Phenomenal Consciousness*. New York: Cambridge University Press.

Robinson, William S. (2006). Knowing Epiphenomena. *Journal of Consciousness Studies*, 13(1-2):85-100.

Robinson, William S. (2010). Your Brain and You: What Neuroscience Means for Us. New York: Goshawk Books.

Robinson, William S. (2012). Phenomenal Realist Physicalism Implies Coherency of Epiphenomenalist Meaning. *Journal of Consciousness Studies*, 19(3-4):145-163.

Robinson, William S. (2013a). Experiencing is not Observing: A Response to Dwayne Moore on Epiphenomenalism and Self-Stultification. *The Review of Philosophy and Psychology*. 4(2):185-192.

Robinson, William S. (2013b). Qualia Realism. In *A Field Guide to the Philosophy of Mind*. http://host.uniroma3.it/progetti/kant/field/qr.htm. Retrieved 2013-06-30. (Date of first publication is unknown to me; citing date of retrieval instead.)

Robinson, William S. (2014). Developing Dualism and Approaching the Hard Problem. *Journal of Consciousness Studies*, 21(1–2):156-182.

Robinson, William S. (2015). Epiphenomenalism. In E. N. Zalta (ed.) *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/epiphenomenalism, accessed 2016-09-20.

Rudd, Anthony. (2000). Phenomenal Judgment and Mental Causation. *Journal of Consciousness Studies*, 7(6):53-66.

Russell, Bertrand. (1917). Knowledge by Acquaintance and Knowledge by Description. In Russell's Mysticism and Logic, Totowa, NJ: Barnes & Noble Books, 1951, pp. 152-167. (References paginated as in the 1951 publication.)

Rutkowski, Chris A. (1984), Geophysical Variables and Human Behavior: XVI.

Some Criticisms. *Perceptual and Motor Skills*, 58:840-842.

Rutkowski, Chris A. (1986). Geophysical Variables and Behavior: XXXIV. Further Comments. *Perceptual and Motor Skills*, 63:18.

Sahu, Satyajit; Ghosh, Subrata; Ghosh, Batu; Aswani, Krishna; Hirata, Kazuto; Fujita, Daisuke and Bandyopadhyay, Anirban. (2013a). Atomic Water Channel Controlling Remarkable Properties of a Single brain Microtubule: Correlating Single Protein to its Supramolecular Assembly. *Biosensors and Bioelectronics*, 47:141–148.

Sahu, Satyajit; Ghosh, Subrata; Hirata, Kazuto; Fujita, Daisuke and Bandyopadhyay, Anirban. (2013b). Multi-Level Memory-Switching Properties of a Single Brain Microtubule. *Applied Physics Letters*. 102:123701.

Saroka, Kevin; Mulligan, Bryce P; Murphy, Todd R. and Persinger, Michael A. (2010). Experimental Elicitation of an Out of Body Experience and Concomitant Cross-Hemispheric Electroencephalographic Coherence. *NeuroQuantology*, 8(4):466-477.

Schrodinger, Erwin. (1958). The Mystery of Sensual Qualities. Chapter 6 of Mind and Matter, in What is Life? with Mind and Matter and Autobiographical Sketches Cambridge University Press, Canto Edition (1992).

Schwartz, Jeffery M; Stapp, Henry P and Beauregard, Mario. (2005) Quantum Physics in Neuroscience and Psychology: A Neurophysical Model of Mind/Brain Interaction. *Phil. Trans. Royal Society, B*. 360 (1458) 1309-27.

Seager, William. (2006). Emergence, Epiphenomenalism and Consciousness. *Journal of Consciousness Studies*. 13(1-2):21-38.

Searle, John R. (1992). The Rediscovery of the Mind. Cambridge, MA: MIT Press.

Searle, John R. (1997). The Mystery of Consciousness. New York: New York Review of Books.

Searle, John R. (1998). How to Study Consciousness Scientifically. in Searle (Ed.) 2002. Consciousness and Language. Cambridge: Cambridge University Press. Online at http://books.google.com/books?id=bvxhV-1Duz8C.

Searle, John R. (1999). Consciousness. Unpublished manuscript. http://socrates.berkeley.edu/~jsearle/Consciousness1.rtf

Searle, John R. (2001). Rationality in Action. Cambridge, MA: MIT Press.

Searle, John R. (2002a). Why I Am Not a Property Dualist. *Journal of Consciousness Studies*. 9(12):57-64.

Searle, John R. (2002b). Consciousness and Language. Cambridge: Cambridge University Press.

Searle, John R. (2004). Mind - A Brief Introduction. New York: Oxford University Press. https://www.scribd.com/doc/262291493/John-Searle-Mind-A-Brief-Introduction retrieved 2015-04-27.

Searle, John R. (2006). Freedom & Neurobiology. New York: Columbia University Press. (Originally given as lectures in Paris in 2001, Searle himself (in 2007b)

cites this work as Searle 2006).

Searle, John R. (2007a). Dualism Revisited. *Journal of Physiology* - Paris 101. 169–178.

Searle, John R. (2007b). Neuroscience, Intentionality and Free Will: Reply to Habermas, *Philosophical Explorations*, 10(1):69-76.

Searle, John R. (2008). The Self as a Problem in Philosophy and Neurobiology. In Philosophy in a New Century: Selected Essays by John R. Searle, Cambridge: Cambridge University Press. pp. 137-151.

Searle, John R. (2013). Can Information Theory Explain Consciousness? *New York Review of Books*. 2013-01-10. http://www.nybooks.com/articles/archives/2013/jan/10/can-information-theory-explain-consciousness/

Simonian, Joseph. (2004). The Paradoxes of Chemical Classification: Why 'Water is $H_2O$' is Not an Identity Statement. *Foundations of Chemistry*, 7(1):49-56.

Seife, Charles. (2006). Decoding the Universe. London:Penguin Books.

Smart, J.J.C. (1959). Sensations and Brain Processes. *The Philosophical Review*, 68(2):141-156.

Smart, J.J.C. (2007). The Mind/Brain Identity Theory. In *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/mind-identity/ retrieved 2016-03-05.

Strawson, Galen. (2010). Mental Reality, 2nd. Ed. Cambridge, MA: MIT Press.

Stapp, Henry P. (2000). Decoherence, Quantum Zeno Effect, and the Efficacy of Mental Effort. http://arxiv.org/pdf/quant-ph/0003065v2 retrieved 2015-04-05.

Stapp, Henry P. (2007). Mindful Universe: Quantum Mechanics and the Participating Observer. Berlin: Springer-Verlag.

Stapp, Henry P. (2008). Philosophy of Mind and the Problem of Free Will in the Light of Quantum Mechanics http://arxiv.org/pdf/0805.0116v1 retrieved 2011-02-08.

Stapp, Henry P. (2009). Physicalism Versus Quantum Mechanics. Chapter 13 in Mind, Matter, and Quantum Mechanics, 3rd Edition, by Henry P. Stapp. Berlin: Springer-Verlag.

Stapp, Henry P. (2010). Free Will. Unpublished manuscript. http://www-physics.lbl.gov/~stapp/FW.pdf retrieved 2015-01-22.

Stoljar, Daniel. (2005). Physicalism and Phenomenal Concepts. *Mind & Language*, 20(5):469–494.

Stroud, John M. (1956). The Fine Structure of Psychological Time. In H. Quastler (ed.), Information Theory in Psychology. Glencoe, IL: Free Press. pp. 174–205.

Swoyer, Chris and Orilia, Francesco. (2011). Properties in *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/properties retrieved 2014-06-15.

't Hooft, Gerard. (2007). The Free-Will Postulate in Quantum Mechanics. http://arxiv.org/pdf/quant-ph/0701097v1. Retrieved 2009-07-11.

Thaheld, Fred H. (2005). Does Consciousness Really Collapse the Wave Function? A Possible Objective Biophysical Resolution of the Measurement Problem. *BioSystems*, 81:113-124.

Thaler L, Arnott SR, Goodale MA. (2011). Neural correlates of natural human echolocation in early and late blind echolocation experts. *PLoS ONE*, 6(5):e20162. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3102086/pdf/pone.0020162.pdf. Retrieved 2015-11-21.

Thomason, Timothy (2013). Projective Perception. Available at: http://works.bepress.com/timothy_thomason/103.

Trefil, James. (1996). 101 Things You Don't Know about Science and No One Else Does Either. New York: Houghton Mifflin Company.

University College London. (2015). Quantum physics problem proved unsolvable. *Science Daily*, 2015-12-09. www.sciencedaily.com/releases/2015/12/151209142727.htm.

Van Gulick, Robert. (2004). So Many Ways of Saying No to Mary. In Peter Ludlow, Yujin Nagasawa and Daniel Stoljar (eds.) There's Something About Mary. Cambridge, MA: MIT Press. pp. 365-405.

VanRullen, Rufin and Koch, Christof. (2003). Is Perception Discrete or Continuous? *TRENDS in Cognitive Sciences*, 7(5):207-213.

VandeWall, Holly. (2007). Why Water Is Not $H_2O$, and Other Critiques of Essentialist Ontology from the Philosophy of Chemistry. *Philosophy of Science*, 74(5).

Vicente, Agustín (2006). On the Causal Completeness of Physics. *International Studies in the Philosophy of Science*, 20(2):149–171.

von Neumann, John. (1955/1932). Mathematical Foundations of Quantum Mechanics. Princeton:Princeton University Press. Translated by Robert T. Beyer.

Voss U; Holzmann R; Tuin I; Hobson A. (2009). Lucid dreaming: a state of consciousness with features of both waking and non-lucid dreaming in *SLEEP*. 32(9):1191-1200. Full text online at http://www.journalsleep.org/ViewAbstract.aspx?pid=27567. Retrieved 2015-04-21.

Wang, Zheng; Solloway, Tyler; Shiffrin, Richard M. and Busemeyer, Jerome R. (2014). Context Effects Produced by Question Orders Reveal Quantum Nature of Human Judgments. *Proceedings of the National Academy of Science*, 111(26):9431-9436.

Wasserman, Ryan. (2004). The Constitution Question. *Nous*, 38(4):693–710.

Watkins, Michael. (1989). The Knowledge Argument Against the Knowledge Argument. *Analysis* 49 (June):158-60.

Weinberg, Steven. (1992). Dreams of a Final Theory. New York: Pantheon Books.

Weinberg, Steven. (2005). Einstein's Mistakes. November, *Physics Today*.

Weingarten, Carol P; Doraiswamy, P. Murali; and Fisher, Matthew PA. (2016). A New Spin on Neural Processing: Quantum Cognition. *Frontiers in Human Neuroscience*. 10:541.

Weisberg, Michael. (2003). Water is *Not* H2O. In Davis Baird, Eric Scerri and Lee McIntyre (eds.), Philosophy of Chemistry. Springer. pp. 337–345.

Wigner, Eugene P. (1962). Remarks on the Mind-Body Question in The Scientist Speculates, I.J. Good, ed., p. 284-302. New York: Basic Books.

Wigner, Eugene P. (1969). Are we machines? *Proceedings of the American Philosophical Society*, 113:95–101.

Wikipedia. (2016-07-03). MIT Blackjack Team. Online at https://en.wikipedia.org/wiki/MIT_Blackjack_Team. Accessed 2016-10-07.

Wikipedia. (2016-09-21). Douglas Mawson. Online at https://en.wikipedia.org/wiki/Douglas_Mawson. Accessed 2016-10-08.

Wikipedia. (2016-11-26). Diamond Color. Online at https://en.wikipedia.org/wiki/Diamond_color. Accessed 2016-12-10.

Wikipedia. (2017-01-05). Methane. Online at https://en.wikipedia.org/wiki/Methane. Accessed 2017-01-07.

Wikipedia. (2017-01-15). Dwarf Planet. Online at https://en.wikipedia.org/wiki/Dwarf_planet. Accessed 2017-01-15.

Wittgenstein, Ludwig. (2001), *Philosophical Investigations*, 3rd Edition. Translated by G.E.M. Anscombe. Malden, MA: Blackwell Publishing. (References to this work are references to paragraph number)

Yablo, Stephen. (1992). Mental Causation, *The Philosophical Review*, 101(2):245-280.