JOURNAL OF PHILOSOPHICAL RESEARCH
VOLUME 28, 2003

THE PROPERTY DUALISM ARGUMENT AGAINST PHYSICALISM

PROFESSOR ANDREW BOTTERELL
SONOMA STATE UNIVERSITY

ABSTRACT: Many contemporary philosophers of mind are concerned to defend a thesis called *a posteriori physicalism*. This thesis has two parts, one metaphysical, and the other epistemological. The metaphysical part of the thesis—the *physicalist* part—is the claim that the psychological nature of the actual world is wholly physical. The epistemolocal part of the thesis—the *a posteriori part*—is the claim that no a priori connection holds between psychological nature and physical nature. Despite its attractiveness, however, a familiar argument alleges that a posteriori physicalism cannot be true. This argument is sometimes called the *Property Dualism Argument Against Physicalism*. In this paper, I consider Stephen White's version of the Property Dualism Argument and argue that it fails. I distinguish two ways in which the argument's crucial notion might be understood, and I argue that on neither way of understanding it is the Property Dualism Argument compelling.

I

Many contemporary philosophers of mind are concerned to defend a thesis that I will call *a posteriori physicalism*. This thesis has two parts, one metaphysical, the other epistemological. The metaphysical part of the thesis—the *physicalist* part—is the claim that the psychological nature of the actual world is wholly physical. So in particular, if there are psychological states, events, processes, or properties, then the physicalist says that they are identical with physical states, events, processes, or properties.[1] The epistemological part of the thesis—the *a posteriori* part—is the claim that no a priori connection holds between psychological nature and physical nature. More precisely:

A posteriori physicalism: For every referring psychological expression ψ there is a referring physical expression ϕ such that $\ulcorner\psi = \phi\urcorner$ is true; and (ii) wherever ψ is a referring psychological expression and ϕ is a referring physical expression and $\ulcorner\psi = \phi\urcorner$ is true, $\ulcorner\psi = \phi\urcorner$ is true and knowable only a posteriori.[2]

A posteriori physicalism is an attractive thesis. Despite its attractiveness, however, a familiar argument alleges that a posteriori physicalism cannot be true. This argument was articulated most famously by J. J. C. Smart in his classic paper "Sensations and Brain Processes," and has received renewed attention and endorsement by Stephen White in his paper "Curse of the Qualia."[3] This argument is sometimes called the *Property Dualism Argument against Physicalism*. My aim in this paper is to consider White's version of the Property Dualism Argument and to argue that it fails.

Although the Property Dualism Argument is not new, there are several reasons for revisiting it. First, although much has been written about the Property Dualism Argument, to my mind nobody has yet produced a convincing refutation of it. And second, the argument—in one form or another—continues to exert an influence on the philosophy of mind, as recent work by Kripke (1980), White (1986), Loar (1990; 1997), Chalmers (1996), Jackson (1998), Levine (1998), and others indicates. It therefore seems to me that a reappraisal of the Property Dualism Argument is due.

II

Since this paper is concerned with physicalism in the philosophy of mind, a criterion that would enable us to clearly distinguish the psychological from the physical might seem to be required. Despite the importance of such a criterion, however, I will leave the distinction between psychological nature and physical nature purposefully vague. There are two reasons for this. First, a rough, intuitive grasp of the distinctions is all that is required to state and understand the Property Dualism Argument. And second, it is notoriously difficult to make these distinctions more precise, and I have no very good suggestions about how this might be done. Some remarks, however, may prove helpful.

By a physical property I have in mind a property that can be had by paradigmatic non-conscious objects, or a property that is constituted by properties that can be had by paradigmatic non-conscious objects. Thus, suppose the predicate $\ulcorner$is $\phi\urcorner$ is part of a *completed physical theory*, i.e., a theory that suffices in principle to explain the nature of all non-conscious objects.[4] Then if P is the property for which the predicate $\ulcorner$is $\phi\urcorner$ stands, P will be a *physical property*. I will call any expression that is part of a completed physical theory a *physical expression*. This metalinguistic definition of "physical property" will play an important role in what follows.

By a psychological property I have in mind a property that can be had only by objects that sometimes think or are sometimes conscious. Paradigmatic examples of such properties are intentional properties—such as the property of believing that pigs fly—and qualitative properties—such as the property of being in pain. I will call any predicate that stands for a psychological property a *psychological predicate*. Any expression that contains a psychological predicate I will call a *psychological expression*.[5]

Although this is obviously rough, it does at least have the virtue of leaving open the possibility that psychological properties might be physical properties. For it may be that psychological properties are constituted by simpler properties that can themselves be had by paradigmatic non-conscious objects. If so, then psychological properties will turn out to be physical in nature.

III

With these preliminaries out of the way, I turn now to Stephen White's version of the Property Dualism Argument.

The Property Dualism Argument is designed to reduce to absurdity a physicalist position that identifies qualitative psychological states, such as Smith's being in pain at *t*, with physical states of the brain, and does so a posteriori. White argues as follows:

> I am assuming, for simplicity, that a person's qualitative state of pain at *t*, say Smith's, is identical with a physical state, say Smith's brain state of type *X* at *t*. Even if this is the case, however, not only do the sense of the expression "Smith's pain at *t*" and the sense of the expression "Smith's brain state *X* at *t*" differ, but the fact that they are coreferential cannot be established on a priori grounds. Thus there must be different properties of Smith's pain (i.e., Smith's brain state *X*) in virtue of which it is the referent of both terms. . . . The general principle is that if two expressions refer to the same object and this fact cannot be established a priori, they do so in virtue of . . . different modes of presentation of that referent. . . . The natural candidates for these modes of presentation are properties. . . .
>
> Since there is no physicalist description that one could plausibly suppose to be coreferential a priori with an expression like "Smith's pain at *t*," no physical property of a pain (i.e., a brain state of type *X*) could provide the route by which it was picked out by such an expression. (White 1986, 91–94)

A general argument is suggested by White's remarks. The core of the argument can be stated as follows:

> (P1) For all expressions α and β, if $\ulcorner \alpha = \beta \urcorner$ is true a posteriori then α and β pick out their common referent in virtue of distinct properties of that referent.

The notion of an expression picking out its referent in virtue of its referent having a particular property is obscure, but let us leave it unanalyzed for the moment. I will return to discussion of it presently.

For the purposes of our reductio, let us focus on some particular psychological expression, such as "Smith's being in pain at *t*." Now, from the thesis of a posteriori physicalism it follows that:

(C1) For every physical expression ϕ, if ⌜Smith's being in pain at $t = \phi$⌝ is true, then ⌜Smith's being in pain at $t = \phi$⌝ is true only a posteriori.

And from (P1) and (C1) it follows that:

(C2) For every physical expression ϕ, "Smith's being in pain at *t*" picks out its referent in virtue of a property that is distinct from the property in virtue of which ϕ picks out its referent.

What follows from (C2)? According to White what follows is that "no physical property of a pain (i.e., a brain state of type *X*) could provide the route by which it was picked out by [an expression like 'Smith's being in pain at *t*']" (White 1986, 94). According to White, that is, if "Smith's being in pain at *t*" picks out its referent in virtue of a property that is distinct from the property in virtue of which any physical expression picks out its referent, then "Smith's being in pain at *t*" picks out its referent in virtue of a non-physical property of its referent.

However, there is a gap in White's reasoning: (C2) does not entail that "Smith's being in pain at *t*" picks out its referent in virtue of a non-physical property of its referent, and so does not entail that a posteriori physicalism is false. For even if the property in virtue of which "Smith's being in pain at *t*" picks out its referent is distinct from the property in virtue of which any physical expression picks out its referent, this is compatible with that property's being a physical property. To see why, suppose that (C2) is true. In particular, suppose that *F* is the property in virtue of which "Smith's being in pain at *t*" picks out its referent. And suppose that no physical expression ϕ picks out its referent in virtue of its referent having *F*. This could entail that *F* is a non-physical property only if a necessary condition for *F*'s being a physical property is that some physical expression picks out its referent in virtue of its referent having *F*. For only if we make this assumption would the fact that *F* is *not* a property in virtue of which any physical expression picks out its referent entail that *F* is a non-physical property. Consequently, White's argument succeeds only if we assume that if *P* is a physical property in virtue of which some psychological expression picks out its referent, then there is some physical expression ϕ that also picks out its referent in virtue of its referent having *P*.[6]

An additional premise is therefore needed. The premise suggested by White's discussion is the following:

(P2) If a psychological expression ψ picks out its referent in virtue of its referent having a physical property P then there is a physical expression ϕ that also picks out its referent in virtue of its referent having P.

To see that the addition of (P2) does result in a valid argument for the conclusion that a posteriori physicalism is false, let us suppose that "Smith's being in pain at *t*" picks out its referent in virtue of its referent having some physical property F. Then by (P2) there is a physical expression ϕ that also picks out its referent in virtue of its referent having F. But if "Smith's being in pain at *t*" and ϕ both pick out their referent in virtue of the same property of that referent, then "Smith's being in pain at *t*" and ϕ will be coreferential a priori, falsifying clause (ii) of the thesis of a posteriori physicalism. So on the assumption that "Smith's being in pain at *t*" picks out its referent in virtue of its referent having a physical property, a posteriori physicalism is false.

On the other hand, suppose that "Smith's being in pain at *t*" picks out its referent in virtue of a property F that is distinct from the property in virtue of which any physical expression picks out its referent. Then no physical expression ϕ will pick out its referent in virtue of its referent having F. But then by (P2) the property in virtue of which "Smith's being in pain at *t*" picks out its referent—i.e., F—will not be a physical property, falsifying clause (i) of the thesis of a posteriori physicalism.

In short, from (P1) and (P2) it follows that:

(C) For all psychological expressions ψ, either ψ picks out its referent in virtue of its referent having a non-physical property, or there is a physical expression with which ψ is coreferential a priori.

And (C) is clearly incompatible with a posteriori physicalism. For (C) entails that anyone who holds that every psychological entity is a physical entity cannot also hold that the relation between psychological entities and physical entities is only a posteriori.

IV

It might be thought that we now have a convincing argument against a posteriori physicalism. However, this conclusion would be premature. For while (P1) and (P2) jointly entail that a posteriori physicalism is false, no a posteriori physicalist is likely to find the Property Dualism Argument compelling. To see why, consider an a posteriori physicalist who accepts (P1). Our a posteriori physicalist presumably thinks that some psychological expressions refer. And because she is a physicalist and accepts (P1), she also thinks that referring expressions pick out their referents via physical properties of those referents. So she is obviously committed to the claim that if ψ is a psychological expression, then ψ picks out its referent in virtue of some physical property of its referent. Hence, any a posteriori physicalist who accepts (P1) is committed to the truth of the antecedent of (P2).

Suppose in addition, however, that our a posteriori physicalist also accepts (P2). Since she accepts (P2), and since she is committed to the truth of the antecedent of (P2), she is also committed to the truth of its consequent. But it is easy to see that the conjunction of (P1) with the consequent of (P2) is incompatible with a posteriori physicalism. For together they entail that ψ and ϕ pick out their common referent in virtue of the same property of that referent. So an a posteriori physicalist who accepts both (P1) and (P2) is committed to the claim that ψ and ϕ are coreferential a priori. But this is simply the denial of clause (ii) of the thesis of a posteriori physicalism.

In short, any a posteriori physicalist who accepts (P1) will *of course* reject (P2), since to accept (P2) in the context of the Property Dualism Argument amounts to accepting the incoherence of a posteriori physicalism. So the question is: is this rejection defensible? Are there any reasons—independent of the role they play in the Property Dualism Argument—that support both (P1) and (P2), and thereby refute a posteriori physicalism? I will argue that there are not.

V

Up to this point my main concern has been to reconstruct a valid argument for the conclusion that a posteriori physicalism is false. This involved supplementing White's version of the Property Dualism Argument with an additional premise. As I pointed out, however, while the addition of an additional premise secures the validity of the resulting argument, no a posteriori physicalist will grant both premises. With this in mind, let me now turn to several influential objections to the Property Dualism Argument. Discussion of these objections will, I hope, make my objection to the argument easier to understand.

Historically, the most influential response to the Property Dualism Argument has been to argue that the argument equivocates on the phrase "non-physical property." J. J. C. Smart, for example, argued that even if an expression like "Smith's being in pain at *t*" picks out its referent in virtue of a non-physical property of its referent, this is compatible with physicalism so long as the property in question is not *objectionably* non-physical. Smart called such unobjectionable non-physical properties *topic-neutral properties*.[7] Nowadays they are more commonly called *functional properties*. Following Smart, let us therefore distinguish two ways in which a property might be said to be non-physical. Let us call a property *weakly non-physical* if having that property does not entail being purely physical in nature. And let us call a property *strongly non-physical* if having that property does entail being non-physical in nature.

The conclusion of the Property Dualism Argument is:

(C) For all psychological expressions ψ, either ψ picks out its referent in virtue of its referent having a non-physical property, or there is a physical expression with which ψ is coreferential a priori.

With the distinction between strongly and weakly non-physical properties in hand, however, (C) no longer entails that a posteriori physicalism is false. For if "non-physical property" means *weakly* non-physical property, then (C) is compatible with physicalism. So it might be thought that the introduction of a distinction between two sorts of non-physical properties shows that the Property Dualism Argument fails.

It seems to me, however, that the core of the Property Dualism Argument remains unaffected by this point. To see why, let us incorporate the notion of a weakly non-physical property into our statement of the Property Dualism Argument, replacing (P2) with (P2*):

> (P2*) If a psychological expression ψ picks out its referent in virtue of its referent having a physical or weakly non-physical property P, then there is a physical or weakly non-physical expression ϕ that also picks out its referent in virtue of its referent having P.

And let us explicitly mention weakly non-physical properties in our definition of a posteriori physicalism:

> *A posteriori physicalism**: (i) For every referring psychological expression ψ there is a referring physical or weakly non-physical expression ϕ such that $\ulcorner\psi = \phi\urcorner$ is true; and (ii) wherever ψ is a referring psychological expression and ϕ is a referring physical or weakly non-physical expression and $\ulcorner\psi = \phi\urcorner$ is true, $\ulcorner\psi = \phi\urcorner$ is true only a posteriori.

Even if these changes are made, however, the conclusion that a posteriori physicalism is false still follows. To see why, let F be the property in virtue of which "Smith's being in pain at t" picks out its referent. Now we know that if F is a physical property, then clause (ii) of the thesis of a posteriori physicalism is false; and we also know that if F is a strongly non-physical property, then clause (i) of the thesis of a posteriori physicalism is false. What follows on the supposition that F is a weakly non-physical property? If F is a weakly non-physical property, then by (P2*) there will be a weakly non-physical expression ϕ that also picks out its referent in virtue of its referent having F. But if "Smith's being in pain at t" and ϕ pick out their referent in virtue of the same weakly non-physical property of that referent, then "Smith's being in pain at t" and ϕ will be coreferential a priori, again falsifying clause (ii) of our revised thesis of posteriori physicalism.

In short, even given a distinction between strongly non-physical and weakly non-physical properties property dualism will be avoided only if every psychological expression is coreferential a priori with a weakly non-physical expression. Some philosophers will happily accept this conclusion.[8] Many, however, will not. Consequently, I think we would do well to look elsewhere for a response to the Property Dualism Argument.

VI

A different response to the Property Dualism Argument takes as its starting point the new theory of reference articulated by Saul Kripke and Hilary Putnam.[9] Rather than distinguishing between different kinds of non-physical properties, this response distinguishes between different ways in which expressions can pick out referents. In doing this, this response rejects the semantic principle underlying the first premise of the Property Dualism Argument. A sophisticated development of this idea can be found in Brian Loar's paper "Phenomenal States."[10]

On Loar's view, two expressions can converge a posteriori on a common referent without picking out that referent in virtue of distinct properties of that referent. This will be the case if one of the expressions refers directly, without the mediation of any property whatsoever. Thus according to Loar, the anti-physicalist conclusion of the Property Dualism Argument can be avoided "if a [qualitative expression or] concept can pick out a physical property directly or essentially, not via a contingent mode of presentation" (1997, 600). In such a case we might say that the qualitative concept is "triggered by" its referent. For lack of a better term, let us call this *the triggering view*.

By way of illustration, suppose the qualitative psychological expression "Smith's being in pain at *t*" and the physical expression "Smith's brain state of type *X* at *t*" are coreferential a posteriori. According to the Property Dualism Argument, it follows that "Smith's being in pain at *t*" and "Smith's brain state of type *X* at *t*" pick out their common referent in virtue of distinct properties of that referent. On the triggering view, however, this is only partly correct: for even if "Smith's brain state of type *X* at *t*" picks out its referent via some property of that referent, this is compatible with "Smith's being in pain at *t*" picking out its referent—i.e., Smith's brain state of type *X* at *t*—directly. Clearly, if "Smith's being in pain at *t*" picks out its referent directly, then *a fortiori* it does not pick out its referent in virtue of any property of that referent, and so does not pick out its referent via a property that is distinct from the property in virtue of which "Smith's brain state of type *X* at *t*" picks out its referent.

The triggering view has much to recommend it as a response to the Property Dualism Argument. However, it is not without problems. The main problem has to do with the concept of triggering itself. The core of the triggering view is the idea that physical states can directly trigger the application of qualitative concepts. But this gives rise to an obvious question, namely, what makes it the case that "Smith's being in pain at *t*," say, is triggered by brain states of type *X* rather than by brain states of type *Y*? What makes it the case, in other words, that "Smith's being in pain at *t*" picks out pain rather than itching? Proponents of the triggering view cannot say that it is because brain states of type *X* are painful that they are picked out by "Smith's being in pain at *t*," since that would be tantamount to saying that "Smith's being in pain at *t*" picks out brain states of type *X* in virtue of those brain states having the

property of being painful. And that is precisely the semantic picture to which the triggering view is intended to be an alternative.

To be sure, this argument is not demonstrative. For example, it might be argued that all that is required for a brain state of type *X* to trigger the application of a qualitative psychological expression like "Smith's being in pain at *t*" is for a law-like relation to hold between "Smith's being in pain at *t*" and brain states of type *X*. Alternatively, it might be argued that the triggering view's claim is not that qualitative psychological expressions pick out their referents without the mediation of any property, but rather that qualitative psychological expressions pick out their referents in virtue of essential properties of those referents. Against this, however, note that this sort of view is often endorsed by philosophers who wish to argue *against* physicalism.[11] Thus, it seems to me that more work needs to be done before the triggering view can be said to constitute a satisfying response to the Property Dualism Argument.

Perhaps such work can be done, and something like the triggering view can be made to work. Other things being equal, however, it would be preferable if our response to the Property Dualism Argument did not rest on contentious semantic assumptions.[12] Consequently, my objection to the Property Dualism Argument will take a somewhat different form. Rather than questioning the semantic principle at work in the Property Dualism Argument, I will argue that even if this semantic principle is granted, the argument still fails.

VII

Before turning to my objection to the Property Dualism Argument, however, I need to say something more about the notion of *an expression picking out its referent in virtue of its referent having a particular property*. In order to facilitate discussion, I will say that if an expression *a* refers to its referent in virtue of its referent having property *P*, then *a introduces P*.[13] White says very little by way of explanation of this notion. Still, given White's references to modes of presentations, two natural interpretations suggest themselves. In brief, "introduce" can be interpreted in a broadly Fregean or a broadly Kripkean manner.[14]

According to the broadly Fregean interpretation of "introduce" I have in mind—the *Fregean interpretation* for short—an expression α is said to introduce a property *P* just in case *P* is the sense of α. For example, suppose "water" is synonymous with the definite description, "the stuff that falls from the sky in the form of rain, fills the oceans and lakes, and is a colorless, odorless liquid." Then on the reasonable assumption that the predicate "is the stuff that falls from the sky in the form of rain, fills the oceans and lakes, and is a colorless, odorless liquid" stands for the property of being the stuff that falls from the sky in the form of rain, fills the oceans and lakes, and is a colorless, odorless liquid, that property will be the sense of "water," and so will be introduced by "water." More generally, according to the Fregean interpretation of

"introduce," an expression α is synonymous with a description D just in case α has as its sense the property for which the predicate ⌜is D⌝ stands.[15]

I turn now to the Kripkean interpretation of "introduce." Kripke (1980) defends a semantic theory according to which proper names and natural kind terms pick out their referents via reference-fixing properties, but without the mediation of a Fregean sense. This suggests another way to make sense of the claim that expressions introduce properties. According to the broadly Kripkean interpretation of "introduce" I have in mind—the *Kripkean interpretation* for short—an expression α is said to introduce a property P just in case P fixes the reference of α. Thus, suppose that the reference of "Bill Clinton" is fixed by the property of being the President of the United States. If so, then "Bill Clinton" will introduce the property of being the President of the United States. However, since according to Kripke fixing the reference of an expression is not the same as giving the meaning of that expression, a property P can fix the reference of an expression α even if P is associated with α only a posteriori, and is not synonymous with α.[16]

In short, the crucial notion of an expression introducing a property admits of two different interpretations, a broadly Fregean interpretation and a broadly Kripkean one. This results in two different versions of the Property Dualism Argument. My argument will be simple. I will argue that if "introduce" is interpreted in a broadly Fregean manner, then there is no reason to think that (P2) is true. I will then argue that if "introduce" is interpreted in a broadly Kripkean manner, (P1) is false. This will suffice to show that on no single interpretation of "introduce" are (P1) and (P2) both true.

VIII

With this in mind, let us turn to the Fregean interpretation of the Property Dualism Argument. Interpreting "introduce" in a broadly Fregean manner gives us the following two interpretations of (P1) and (P2):

> *(P1-Fregean)*: For all expressions α and β, if ⌜α = β⌝ is true only a posteriori then for some predicate F, α is synonymous with ⌜the F⌝, and for some predicate G, β is synonymous with ⌜the G⌝, and ⌜the property of being F = the property of being G⌝ is false.

> *(P2-Fregean)*: If a psychological expression ψ picks out its referent in virtue of ψ having a physical property P as its sense, then there is a physical expression ϕ such that ϕ also picks out its referent in virtue of having P as its sense.

Let us focus our attention on (P2-Fregean). (P2-Fregean) is not without advocates. Brian Loar, for example, appeals to a version of (P2-Fregean) when he argues that if a physical property P were the sense of some psychological expression ψ, "there would be an a priori connection between that [psychological] term [ψ] and some physical term, viz., one that more explicitly

expresses that sense [i.e., *P*]" (1990, 84, n. 5). Loar is here stating what he takes to follow from the adoption of (P1-Fregean). However, since Loar does not provide an explicit argument for this claim, it stands to reason that he thinks that it is obvious that a physicalist who endorses (P1-Fregean) is committed to thinking that if a physical property *P* is the sense of some psychological expression ψ, then there is some physical expression ϕ that also has *P* as its sense. But is this obvious?

If the argument is to have any force, proponents of the Fregean version of the Property Dualism Argument need to provide some reason for thinking that (P2-Fregean) is true. It seems to me that the best way to establish the truth of (P2-Fregean) is to provide an argument for the conclusion that every physical property is the sense of some physical expression or other. What I will argue, however, is that there is no compelling argument in favor of this thesis. I will do this in two stages. First, I will consider a natural argument for the conclusion that every physical property is the sense of some physical expression, and provide some reasons for thinking that it is misguided. Next, I will consider a more complicated argument for the same conclusion, and argue against it. Since this strategy is piece-meal, it will no doubt be objected that I have overlooked another, more promising, line of argument in favor of (P2-Fregean). Fair enough. All I am trying to do is show that we are owed an argument for (P2-Fregean), and that it is not obvious that such an argument is forthcoming.

IX

The first argument in favor of (P2-Fregean) that I will discuss proceeds as follows: suppose *P* is a physical property, and suppose ⌜is *F*⌝ is the physical predicate that stands for *P*. Then given the existence of the predicate ⌜is *F*⌝ prevent us from forming the expression ⌜the *x* such that *Fx*⌝. And on the assumption that an expression of the form ⌜the *x* such that *Fx*⌝ has as its sense the property for which the predicate ⌜is *F*⌝ stands, every physical property will be the sense of some physical expression or other, and so will be introduced by some physical expression or other. So (P2-Fregean) is true.

This argument is straightforward. However, consider the assumption that the sense of a definite description ⌜the *x* such that *Fx*⌝ is the property for which the predicate ⌜is *F*⌝ stands. Given this assumption it would seem to follow that if two predicates stand for the same property then it is a priori that the corresponding descriptions are coreferential, since those descriptions will have the same sense. But there are good reasons for thinking that this is false. For example, while it is plausible to suppose that the predicate "is water" and the predicate "is H_2O" stand for the same property, it is arguable that "the *x* such that *x* is water = the *x* such that *x* is H_2O" is not true a priori. Hence, it is arguable that the above argument for (P2-Fregean) rests on a faulty premise.

It will perhaps be objected that all this shows is that the sense of the definite description "the *x* such that *x* is water" cannot be the property of being

H_2O but must instead be some other property, say the property of being a colorless, odorless liquid that falls from the sky in the form of rain, and so on. However, this misses the point. For so long as it is possible for two predicates to be coextensive only a posteriori, the conclusion that the corresponding *descriptions* will be coreferential only a posteriori will still follow. And this possibility is all that is needed in order to show that the crucial premise of the above argument is problematic.[17]

X

Thus far I have been concerned to argue that a natural argument for (P2-Fregean) can be resisted. But might (P2-Fregean) be true nonetheless? It might be thought that the following considerations show that it must be.

Suppose P is a physical property that is the sense of some psychological expression ψ. Then either P is the sense of an expression of some completed physical theory L or it is not. If P is the sense of an expression of L, then (P2-Fregean) is true. If P is *not* the sense of an expression of L then, given the existence of a physical theory like L, there is no reason why there couldn't be another physical theory, L^*, which is just like L except for containing an additional physical expression that has P as its sense. To see that the existence of a theory like L^* is not in doubt, we need only consider the following. Take L. Add to it a new physical expression ϕ^*. Call the new theory L^*. Stipulate that ϕ^* has P as its sense.[18] Then, since by hypothesis P is the sense of a psychological expression and since by stipulation P is the sense of a physical expression, namely ϕ^*, it would appear that if P is the sense of a psychological expression then it is also the sense of a physical expression, in which case (P2-Fregean) is true.

While there are a number of things that might be said about the above line of argument, I think that a posteriori physicalists should not be persuaded by it. Instead, a posteriori physicalists should deny that L^* is relevant to the truth or falsity of (P2-Fregean). This might seem odd. For if L is an acceptable physical theory then there is no obvious reason for thinking that L^* is not an acceptable physical theory. After all, L^* contains only physical expressions that refer to physical entities by having physical properties as senses. And L^* differs from L only in containing an additional expression that has as its sense a physical property, which was not the sense of any expression of L. But it is hard to see how this fact could entail that L^* is not an acceptable physical theory. And if L^* is an acceptable physical theory, then surely it is relevant to the truth or falsity of (P2-Fregean).

Nonetheless, I think that a posteriori physicalists have good reason to object to the relevance of L^*. To see why, recall the conclusion of the Property Dualism Argument: unless there are physical expressions with which psychological expressions are coreferential a priori, physicalists will be forced to acknowledge the existence of non-physical properties. The above argument,

however, only establishes that if a physical property *P* is the sense of a psychological expression ψ then there *could be* a physical expression that has *P* as its sense. However, on one way of understanding this claim, it is irrelevant to the question whether a posteriori physicalism is true; and on the other way of understanding it, it trivializes the debate between a posteriori and a priori physicalists. Let me explain.

The first way of understanding the claim is that for all the a posteriori physicalist has argued there *could be* physical expressions with which psychological expressions are coreferential a priori. However, a posteriori physicalists can surely agree with this and still deny that the proponent of the Property Dualism Argument has thereby established the incoherence of a posteriori physicalism. Of course there *could be* physical expressions with which psychological expressions are coreferential a priori; the question, however, is whether, if physicalism is true, there *must be* physical expressions with which psychological expressions are coreferential a priori. And the above argument for (P2-Fregean) gives us no reason to think that this is indeed the case.

The other way of understanding the claim is that whenever a physical property *P* is the sense of a psychological expression, we can always create a new physical expression and stipulate that the newly created physical expression has *P* as its sense. However, if this is how the claim is to be understood, then it seems clear that the Property Dualism Argument is rendered completely uninteresting. For while it is true that new expressions can always be introduced into a physical theory, and can be stipulated to be synonymous with existing psychological expressions, it is hard to see any threat to the coherence of a posteriori physicalism here.

By way of illustration, consider what Alonzo Church has called the "Principle of Tolerance." The Principle of Tolerance asserts that "everyone is at liberty to build his own form of language as he will" (Church 1954, 160). So in particular, it might be argued that according to the Principle of Tolerance *physicalists* are at liberty to build their own form of physical language. Moreover, since one way to build a physical language is to build a language in which every psychological expression ψ is coreferential a priori with some physical expression ϕ, it might be thought that something like the Principle of Tolerance supports (P2-Fregean), and counts against a posteriori physicalism.

Now it seems to me that if something like Church's Principle of Tolerance is adopted then the truth of (P2-Fregean) can be established. But it also seems to me that establishing the truth of (P2-Fregean) in this manner makes it unclear what is at issue between a posteriori and a priori physicalists. A posteriori physicalists, recall, claim that no a priori connection needs to hold between physical nature and psychological nature in order for physicalism to be true; a priori physicalists deny this. However, it is unclear that the possibility that for any psychological expression ψ we can always *create* a physical expression ϕ with which ψ is coreferential a priori shows that a posteriori physicalism is incoherent. For if it did, then it would seem that we could equally well

argue against a posteriori meteorology by noting that for any meteorological expression *M* we could always create a physical expression with which *M* is coreferential a priori. But this possibility hardly shows that a posteriori meteorology is incoherent.

My aim in these two sections has been to argue that it is not obvious that there is a compelling argument in favor of (P2-Fregean). For either such an argument entails that obvious statements of a posteriori identity are in fact a priori, or the argument is irrelevant to the debate between a priori and a posteriori physicalists.[19]

XI

Let us now consider whether the Kripkean version of the Property Dualism Argument fares any better. If "introduce" is interpreted in a broadly Kripkean manner, we get the following interpretations of (P1) and (P2):

> *(P1-Kripkean)*: For all expressions α and β, if ⌜$\alpha = \beta$⌝ is true a posteriori then α and β pick out their common referent in virtue of distinct reference-fixing properties *F* and *G*, α picking out its referent in virtue of *F* and β picking out its referent in virtue of *G*.

> *(P2-Kripkean)*: If a psychological expression ψ picks out its referent in virtue of a reference-fixing physical property *P*, then there is a physical expression ϕ such that ϕ also picks out its referent in virtue of *P*.

Let us turn to consideration of (P1-Kripkean).

(P1-Kripkean) states that for all expressions α and β, if ⌜$\alpha = \beta$⌝ is true a posteriori then α and β pick out their common referent in virtue of distinct reference-fixing properties *F* and *G*, α picking out its referent in virtue of *F* and β picking out its referent in virtue of *G*. By contraposition we have that for all expressions α and β, if it is not the case that α and β pick out their common referent in virtue of distinct reference-fixing properties of that referent, then it is not the case that ⌜$\alpha = \beta$⌝ is true a posteriori. More precisely: if α and β pick out their common referent in virtue of the same reference-fixing property of that referent, then ⌜$\alpha = \beta$⌝ is true a priori. What I wish to argue is that this claim is false if "introduce" is interpreted in a broadly Kripkean manner. This will suffice to show that (P1-Kripkean) is false, and hence, that the Kripkean version of the Property Dualism Argument is unsound.

In general, it is a mistake to conclude from the fact that the reference of two expressions is fixed by the same property that those two expressions are coreferential a priori. To illustrate this point, consider a true a posteriori identity statement involving two definite descriptions ⌜the x such that *Fx*⌝ and ⌜the *x* such that *Gx*⌝ where the predicates ⌜is *F*⌝ and ⌜is *G*⌝ stand for the same property. Then, according to the Kripkean interpretation of "introduce," ⌜the *x* such that *Fx*⌝ will introduce as a reference-fixing property the property for which the predicate ⌜is *F*⌝ stands, and ⌜the *x* such that *Gx*⌝ will introduce as a

reference-fixing property the property for which the predicate ⌜is G⌝ stands. But by hypothesis the property for which the predicate ⌜is F⌝ stands is the *same* property as the property for which the predicate ⌜is G⌝ stands. Thus the definite descriptions ⌜the x such that Fx⌝ and ⌜the x such that Gx⌝, although coreferential only a posteriori, introduce the same reference-fixing property, and so constitute a counter-example to the claim that two expressions that introduce the same reference-fixing property must be coreferential a priori.

How does this apply to the particular case of the psychological? Since we are granting (P2-Kripkean), we know that if the reference of a psychological expression such as "Smith's being in pain at t" is fixed by a physical property P there will be a physical expression, such as ⌜Smith's being F⌝—where the physical predicate ⌜is F⌝ stands for P—whose reference is also fixed by P. Does it follow from this that "Smith's being in pain at t" is coreferential a priori with an expression like ⌜Smith's being F⌝? I do not see that it does.

To see why, we need only suppose–along with the a posteriori physicalist—that "the property of being in pain" and ⌜the property of being F⌝—where ⌜is F⌝ stands for a physical property—are coreferential only a posteriori. If this is supposed to be the case, then although "Smith's being in pain at t" and ⌜Smith's being F⌝ will introduce the same reference-fixing property, they will not be coreferential a priori. In consequence, there would appear to be nothing wrong with holding (i) that a physical property P is introduced as a reference-fixing property by a psychological expression ψ; (ii) that P is also introduced as a reference-fixing property by a physical expression ϕ; and (iii) that ψ and ϕ are coreferential only a posteriori. In a way this should come as no surprise. For as I noted above, the core of the Kripkean interpretation is that there need be no a priori connection between an expression α and the property introduced by α as a reference-fixing property. And once the lack of such a connection is acknowledged the inference from the claim that two expressions introduce the same reference-fixing property to the conclusion that those expressions must be coreferential a priori is blocked.

Of course, it might be wondered how somebody could competently use and understand an expression like "Smith's being in pain at t," and also competently use and understand an expression like ⌜Smith's being F⌝, and yet not recognize that the two expressions are coreferential. To wonder this, however, is to fail to appreciate the picture of reference underlying the Kripkean interpretation of "introduce." For since according to the Kripkean picture understanding an expression α does not require knowing which property fixes the reference of α, somebody can use the expression "Smith's being in pain at t" to pick out Smith's brain state of type X at t even if that person does not know how that expression picks out the state it does.

XII

One final observation. It is worth noting that one could combine the Kripkean interpretation and the Fregean interpretation to produce a hybrid view. Thus, it might be argued that what accounts for the distinctness of our concepts of qualitative psychological states and our concepts of physical states is that qualitative psychological concepts pick out their referents via reference-fixing properties, while physical concepts pick out their referents via properties that function as senses. So, suppose the physical concept ϕ and the psychological concept ψ are coreferential, that they both pick out their common referent in virtue of that referent having physical property P, but that P is associated with ψ as a reference-fixing property while P is associated with ϕ as its sense. Then it could very well be the case that ϕ and ψ are coreferential only a posteriori, even though they pick out their referent via the same property of that referent.

The hybrid view may very well be what Loar (1990; 1997) is advocating; it is also a reasonable way to interpret Levine (1998). Levine suggests that one way to respond to the Property Dualism Argument is to distinguish two ways in which properties can function as modes of presentation. Levine proposes two kinds of modes of presentation: ascriptive and non-ascriptive. He explains them as follows:

> [a]n ascriptive mode [of presentation] is one that involves the ascription of properties to the referent, and it's (at least partly) by virtue of its instantiation of these properties that the object . . . is the referent. A non-ascriptive mode [of presentation] is one that reaches its target, establishes a referential relation, by some other method. The object isn't referred to by virtue of its satisfaction of any conditions explicitly represented in the mode of presentation, but rather by its standing in some particular relation to the representation. (Levine 1998, 457)

On the assumption that the properties that function as ascriptive modes of presentation are senses, and that the properties that function as non-ascriptive modes of presentation are mere reference-fixers, the hybrid view and Levine's view turn out to be identical.[20]

I mention this to emphasize that the Fregean interpretation of "introduce" and the Kripkean interpretation of "introduce" are not incompatible. On the contrary, the idea that qualitative psychological expressions might pick out their referents in a manner that is distinct from the manner in which physical expressions pick out their referents is both intuitively plausible and theoretically satisfying.

XIII

In conclusion, I have argued that the Property Dualism Argument provides no reason for physicalists to suppose that physicalism must take an a priori form. My strategy involved distinguishing two ways in which the argument's

crucial notion of an expression introducing a property might be interpreted. I first argued that if "introduce" is interpreted in a broadly Fregean manner, then there is no reason to think that (P2-Fregean) is true. I next argued that if "introduce" is interpreted in a broadly Kripkean manner, then there is good reason to think that (P1-Kripkean) is false. Because there is no univocal interpretation on which both (P1) and (P2) are true, I therefore concluded that the Property Dualism Argument fails. I further suggested that my objection is preferable to the objection offered by Smart, since it succeeds in defending a posteriori physicalism, and that my objection is preferable to the objection I called the "triggering view," since it does not require any controversial semantic assumptions.

As I noted, however, a problem with my objection to the Fregean interpretation of the Property Dualism Argument is that it is piece-meal in nature. For while I criticized two arguments in favor of (P2-Fregean), I made no attempt to argue that there could not be another, more compelling, argument for (P2-Fregean). Because of this, proponents of the Property Dualism Argument may reasonably feel that I have overlooked considerations that count in favor of the argument. Moreover, even if the Property Dualism Argument is unsuccessful, this does not mean that there might not be other reasons for thinking that a posteriori physicalism is false. I concede both of these points. Nonetheless, it seems to me that the observation that the Property Dualism Argument is unsuccessful remains important. For many philosophers have employed versions of the Property Dualism Argument to argue that psychological expressions must be coreferential a priori with non-psychological expressions if physicalism is to be true. I think this is a mistake. The conclusion of this paper should lead those philosophers who have been convinced by the Property Dualism Argument to rethink their reasons for holding physicalism in an a priori form.[21]

ENDNOTES

1. This statement of physicalism is clearly stronger than the view that the psychological nature of the actual world is merely necessitated by the physical nature of the actual world. Although physicalism is typically formulated as a supervenience thesis, for the purposes of this paper I will be concerned with identity thesis versions of physicalism only. There are several reasons for this. First, the argument with which I will be concerned has as its target identity thesis versions of physicalism. And second, supervenience thesis versions of physicalism are often taken to carry with them problems of their own. Nonetheless, I believe that the conclusions I reach carry over to weaker versions of physicalism; for arguments to this effect, see Block and Stalnaker (1999) and Byrne (1999).

2. A posteriori physicalism is thus to be contrasted with a priori physicalism, here understood to be the view that (i) for every referring psychological expression ψ there is a referring physical expression ϕ such that $\ulcorner\psi = \phi\urcorner$ is true; and (ii) wherever ψ is a referring psychological expression and ϕ is a referring physical expression and $\ulcorner\psi = \phi\urcorner$ is true, $\ulcorner\psi = \phi\urcorner$ is knowable a priori. For discussion of the merits of a priori physicalism, see Lewis (1994), Chalmers (1996), Jackson (1998), Byrne (1999), Block and Stalnaker (1999), and Stoljar (2000).

3. See also Kripke (1980), Loar (1990; 1997), Chalmers (1996), and Levine (1998).

4. See Block (1978), Lewis (1983), and Jackson (1998) for definitions of "physical property" along similar lines.

5. Given my definitions of "physical property" and "psychological property," it might seem that the thesis of a posteriori physicalism is trivially false. For consider some physical property P, and suppose that ⌜is F⌝ is the physical predicate which stands for P. Then the expression "the x such that x is F or x is in pain" will count as a physical expression—in light of the fact that it contains the physical predicate ⌜is F⌝—and it will also count as a psychological expression—in light of the fact that it contains the psychological predicate "is in pain." Moreover, since every expression is coreferential a priori with itself, and since any psychological expression can be turned into a physical expression by disjoining it with some physical expression, it might seem that every psychological expression *is* coreferential a priori with some physical expression.

I allow that given this very loose notion of a physical expression, it is true that every psychological expression is coreferential a priori with some physical expression; what I deny is that this shows that the thesis of a posteriori physicalism is incoherent. This sort of objection will be dealt with in more detail below.

6. Or, equivalently, if we assume that if there is no physical expression ϕ that picks out its referent in virtue of its referent having P, then P is not a physical property.

7. See Smart (1959), Lewis (1966), and Armstrong (1968).

8. See, for example, Lewis (1966; 1994), Armstrong (1968), Shoemaker (1984), and White (1986).

9. For classic expositions of the new theory of reference see Kripke (1980) and Putnam (1975).

10. See also Levine (1998) for a discussion of this sort of view.

11. See, for example, Kripke (1980).

12. I am not claiming that the new theory of reference is false. I am merely arguing that, as employed by Loar, it leads to problems when applied to the special case of the psychological.

13. This terminology is due to Loar (1990).

14. I think it is reasonably clear that White himself has the Fregean interpretation of "introduce" in mind.

15. This is not intended as an interpretation of Frege's views on the relation between senses and definite descriptions.

16. Kripke also argues that reference-fixing properties can sometimes be associated with expressions a priori. I will return to this issue below.

17. Perhaps proponents of the Fregean version of the Property Dualism Argument will insist that whenever two predicates are coextensive they are coextensive a priori and hence, that there cannot be cases of predicates that are coextensive only a posteriori. However, while this would certainly rebut the argument just presented, it is clearly a further claim, and as such stands in need of justification. Moreover, since such a claim will certainly be rejected by a posteriori physicalists, it seems to me that proponents of the Fregean version

of the Property Dualism Argument cannot reject this argument in this manner without begging the question against a posteriori physicalists.

18. It might be objected that since there must be some constraints on what counts as an acceptable physical expression, the proponent of the Property Dualism Argument cannot simply *stipulate* that ϕ^* introduces P as a sense. I am sympathetic to this objection, and will discuss it in detail below.

19. As an anonymous referee pointed out to me, my objection to (P2-Fregean) bears certain affinities to Richard Boyd's (1980) response to Kripke's (1980) anti-physicalist argument. There Boyd argues that there could be an explanation of how a sentence like "pain = c-fibres firing" could appear contingent from the right-hand side, as it were, even if not from the left-hand side. This involves rejecting the idea that we pick out natural kinds only via essential properties of those kinds. I do think that there are reasons for rejecting this idea, and I also think that it is important to consider the role this idea plays in various arguments for and against physicalism. Unfortunately, due to limitations of space I cannot comment on this issue in more detail here. For some idea of how Boyd's strategy might be used to respond to other anti-physicalist arguments, see Botterell (2001).

20. On the other hand, since a non-ascriptive mode of presentation is supposed to involve a relation—such as causal covariation or triggering—that need not be cognitively accessible, it might be doubted whether a reference-fixing property could function as a non-ascriptive mode of presentation. However, since it does seem to me that a reference-fixing property could fix the reference of an expression even if it is not known to do so, I am inclined to think that the above assumption is defensible. I am indebted to an anonymous referee for raising this issue.

21. For comments on previous drafts I am indebted to Ned Block, Alex Byrne, Lenny Clapp, Judith Feldmann, Ned Hall, Robert Stalnaker, Daniel Stoljar, Judith Jarvis Thomson, and Stephen White. I would also like to thank an anonymous referee for this journal for comments that improved the paper. The writing of this paper was funded in part by a grant from the Social Sciences and Humanities Research Council of Canada; I am grateful for their support.

REFERENCES

Armstrong, D. 1968. *A Materialist Theory of the Mind*. London: Routledge.

Block, N., ed. 1980. *Readings in the Philosophy of Psychology,* vol. 1. Cambridge, Mass.: Harvard University Press.

———— . 1978. "Troubles with Functionalism." In Block (1980), 268–305.

———— and R. Stalnaker. 1999. "Conceptual Analysis, Dualism, and the Explanatory Gap." *The Philosophical Review* 108: 1–46.

Botterell, A. 2001. "Conceiving What Is Not There." *Journal of Consciousness Studies* 8 (August), 21–42.

Boyd. 1980. "Materialism without Reductionism: What Physicalism Does Not Entail." In Block (1980), 67–106.

Byrne, A. 1999. "Cosmic Hermeneutics." *Philosophical Perspectives* 13: 347–383.

Chalmers, D. 1996. *The Conscious Mind*. Oxford: Oxford University Press.

Church, A. 1954. "Intensional Isomorphism and Identity of Belief." *Philosophical Studies* 5: 65–73. Reprinted in *Propositions and Attitudes*, 159–168. Edited by N. Salmon and S. Soames. Oxford: Oxford University Press, 1988.

Jackson, F. 1998. *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Kripke, S. 1980. *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

Levine, J. 1998. "Conceivability and the Metaphysics of Mind." *Nous* 32: 449–480.

Lewis, D. 1966. "An Argument for the Identity Theory." *Journal of Philosophy* 63: 17–25. Reprinted in Lewis, *Philosophical Papers,* vol. 1, 99–107. Oxford: Oxford University Press, 1983.

————. 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61: 343–77.

————. 1994. "Reduction of Mind." In *A Companion to the Philosophy of Mind*, 412–431. Edited by S. Guttenplan. Oxford: Blackwell.

Loar, B. 1990. "Phenomenal States" (original version). In *Philosophical Perspectives*, vol. 4., 81–108. Edited by J. Tomberlin. Atascadero, Calif.: Ridgeview Press.

————. 1997. "Phenomenal States" (revised version). In *The Nature of Consciousness*, 597–616. Edited by N. Block, O. Flanagan, and G. Guzeldere. Cambridge, Mass.: MIT Press.

Lycan, W. 1987. *Consciousness*. Cambridge, Mass.: MIT Press.

Papineau, D. 1993. *Philosophical Naturalism*. Cambridge, Mass.: Blackwell.

Putnam, H. 1975. "The Meaning of 'Meaning.'" In *Mind, Language, and Reality: Philosophical Papers*, vol. 2, 304–324. Edited by Putnam. Cambridge: Cambridge University Press.

Shoemaker, S. 1984. *Identity, Cause, and Mind*. Cambridge: Cambridge University Press.

Smart, J. J. C. 1959. "Sensations and Brain Processes." *The Philosophical Review* 68: 141–156. Reprinted in *Materialism and the Mind-Body Problem*, 53–66. Edited by D. Rosenthal. Indianapolis: Hackett, 1987.

Stoljar, D. 2000. "Physicalism and the Necessary A Posteriori." *The Journal of Philosophy* 97: 33–54.

Tye, M. 1995. *Ten Problems of Consciousness*. Cambridge, Mass.: MIT Press.

White, S. 1986. "Curse of the Qualia." *Synthese* 68: 333–368. Reprinted in *The Unity of the Self*, 75-101. Edited by S. White.