# **Theory Dualism and the Metalogic of Mind-Body Problems**

T. Parent

## *1. Introduction*

What is philosophy good for? The importance of practical philosophy may be obvious enough—but what of theoretical philosophy (a.k.a. "impractical philosophy")? Hofweber (2009) raises this in an acute way for metaphysics, and 'metaphysics' here could just as well be construed to include all of theoretical philosophy:

> the greatest threat to metaphysics as a philosophical discipline…[is] that the questions that metaphysics tries to answer have long been answered in other parts of inquiry, ones that have much greater authority. And if they haven't been answered yet then one should not look to philosophy for an answer. What metaphysics tries to do has been or will be done by the sciences. There is nothing left to do for philosophy (p. 260)

This sentiment plausibly explains why (e.g.) contemporary philosophy of mind looks more and more like cognitive psychology. It is as if practitioners are coming to agree that *there is no* distinctly philosophical work to be done. Now I am hardly one to frown on empirical inquiry. And generally, it is crucial for all philosophers to stay abreast of the latest developments in science—that should go without saying. However, is there anything distinctly philosophical (apart from practical philosophy) that advances our knowledge and understanding? Increasingly, it seems like the answer is 'no'.

However, I suspect there is a false presupposition in such cynicism—namely, that theoretical philosophy is defined mostly by its *topic* or subject-matter. If that view is adopted, then these sub-disciplines indeed look outmoded. For the sciences now cover many of the same topics. As a contrast to the "topical view" of philosophy, however, recall the following lines from the *Tractatus* (my italics):

> 4.112 Philosophy aims at the logical clarification of thoughts.
> *Philosophy is not a body of doctrine but an activity...*
> Philosophy does not result in 'philosophical propositions', but rather in the clarification of propositions.
> Without philosophy thoughts are, as it were, cloudy and indistinct: its task is to make them clear and to give them sharp boundaries.

Admittedly, it is unclear what it is to be "clear." Moreover, the Tractarian view can be questioned on certain points (e.g., 4.113). However, the "activity" conception of philosophy, in contrast to a topical conception, strikes me as getting something right. *Some* philosophical puzzles, at least, remain confusions about the use of our own terms and concepts. In such cases, then, the philosopher might contribute to our collective knowledge and understanding, by removing such confusions as they arise.[1]

Note that philosophy in this vein is not descriptive. Hence, it is not an attempt to do sociolinguistics or cognitive psychology from the armchair.[2] It is rather a prescriptive affair; it aims to remove confusion by enforcing distinctions, imposing precision, and generally clarifying what was otherwise ambiguous, vague, or obscure. Philosophical problems in this vein might be seen as problems in "applied logic," [3] and are rectified in a process that Quine (1960) called *regimentation*.[4]

Since regimentation is not a descriptive program, it does not presume that one can discover, just from the armchair, deep metaphysical truths. (That is more characteristic of "neo-scholastic metaphysics."[5]) Rather, the idea is that some philosophical issues do not need to be resolved so as much as dissolved, e.g., by rooting out some subtle equivocation or false presupposition in the use of a term/concept. Or in some cases, a philosophical puzzle is to be "explained away" as crossing the limits of what can be consistently said (cf. *Tractatus* 4.114). (The semantic paradoxes are the clearest examples of this sort of thing.)

In what follows, I defend the method of regimentation by example. The paper can thus be seen as addressing Hofweber's anxiety by defending (what he loadedly calls) "unambitious metaphysics." For Hofweber, regimentation apparently would count as "unambitious," since it is not concerned to unearth new facts about mind-independent reality. Instead, it only "works out the consequences" of a theory (p. 264). As the present paper illustrates, however, "working out the consequences" means more than just deriving corollaries via trivial inference-rules. Prior to any deriving, the theories in question need to be regimented, and quite a bit hangs on how one chooses to regiment.

I mentioned that philosophy seems particularly outmoded in the theory of mind. Accordingly, the paper focuses on mind-body problems to illustrate the benefits of regimentation. In the main, my remarks concern a problem about after-images, though they are eventually extended to representational or intentional states in general. The conclusion is that intentional states exist, though not in the standard, actualist sense of 'exist'. Indeed, if 'exist' were used in the standard way, then the view here would be a kind of eliminativism.

[1] Distinctions like empirical/non-empirical, discovery/clarity, and individuation by topic/activity I take to be fuzzy. But though sharp distinctions are preferable, fuzzy distinctions can remain of some use. (E.g., it is informative to be told that no swans are blue, even though the color spectrum forms a continuum where no sharp boundaries exist.)

[2] In Frege, there is a thesis which superficially suggests otherwise—viz., he suggests that his is an inquiry into the "structure of thought." But given his anti-psychologism, "thought" for Frege is not a psychological object but rather an *ideal* object. We need not delve into the metaphysical questions about such objects. The point is just that even Frege understood his investigations as having a fundamentally normative orientation.

[3] If some philosophy is a kind of applied logic, it is only natural that it is more activity than doctrine. (Undergrads are loathed to discover that a logic course cannot be passed just by memorizing a set of doctrine.)

[4] On regimentation as prescriptive, see especially Azzouni (2006), chs. 4 and 5. Also, even though I have some serious disagreements with him, see Rosenberg's (1998) "chapter zero."

[5] The term is from Ladyman et al. (2007).

However, since intentional states are said to "exist" in some sense, the view also has an oddly dualist appearance. The two senses of 'exist' can suggest that reality divides, as if there were two fundamental ontological kinds or two "ways of being." Yet I ultimately interpret the situation as not as vindicating a *metaphysical* dualism, but rather as reflecting a dualism of *theory*. Roughly, we can have a theory of mind-independent objects, and a theory of mind-dependent objects. But there may be principled reasons why we cannot integrate these two into a consistent whole. One is just that—as a matter of logical necessity—nothing can be both mind-independent and mind-dependent. Another argument is more complicated, having to do with the heterological paradox and Russell's vicious circle principle.

Such acquiescence to mind-body problems has been called "mysterianism," with McGinn (1989) being the most familiar example. However, unlike McGinn, the present view is not suggesting that these problems could be resolved, if only we were smart enough, or if only we had the right conceptual resources. It is a more thorough-going mysterianism; the idea is that not even God could resolve mind-body problems. For there are logical inconsistencies which arise in the attempt to represent one's representations, as part of the world represented by those same representations.[6]

## *2. The Place-Smart Argument*

The paper was initially drafted to oppose an especially powerful argument for dualism. Indeed, Dennett (2013) declares it to be "the shortest, sweetest, and actually in the end most convincing argument for dualism I know" (12min, 38sec). (It is thus curious why the argument is not discussed more in the literature.) As far as I can tell, the argument was first raised in U.T. Place's classic (1956), though Place himself opposed it. (Jack Smart 1959 also took up the issue, yet responded in much the same way.)

The argument is as follows. Consider a green after-image in your visual field, one that results from (say) flash photography. Then, there is an obvious sense in which the following is true:

(1) There is a green thing.

The green color-patch I take to be a paradigm instance of a "quale." (On my use of this term, see notes 9 and 10.) Now if the thing is a mere after-image, it is uncontentious that:

[6] I recently discovered another "in principle" mysterianism from Molyneux (2011), who also draws upon metalogical concepts. But Molyneux's argument strikes me as not capitalizing on a real problem. His idea is that to identify a mental property *x* and a physical property, one must identify each property of *x* with a property of *y*. But to do that, one needs to identify each property of a property of *x* with a 3rd-order property of *y*, and thus a regress. Yet, as Molyneux seems to concede, this would be a problem in *any* attempt at identification: "the marvel is perhaps not that there are entities such as mind and body that cannot be identified in a finite number of steps, but that there are entities like Hesperus and Phosphorus that can" (p. 227). But since it is *not* a problem in one case, I am unconvinced it is a problem in the other. (For my part, Molyneux's argument more plausibly suggests that identifying *x* and *y* is something we do without explicitly identifying every *n*-order property of *x* with an *n*-order property of *y*. We human beings have only a finite time, and our identifications normally have the status of hypotheses open to empirical revision.)

(2) The green thing is not located outside the skull.

But on reflection, it also seems true that:

(3) The green thing is not inside the skull.

After all, if we were to crack open your skull, we wouldn't find anything green in there. (If we did, you'd be facing some disturbing medical news.) Further, it is plausible to say that:

(4) If both (2) and (3) are true, then the green thing is not in physical space.

And this, conjoined with (1)-(3), entails:

(5) There is a thing that is not in physical space.

More, if we think of occupying the spatial order as the minimum requirement for the entity being physical, it follows:

(6) There is a nonphysical thing.

Call this argument the "Place-Smart argument." I'm inclined to agree with Dennett that the Place-Smart argument is the most compelling argument for dualism there is.

Even so, a physicalist should deny that *everything* physical has a spatial location. There are counterexamples from physics itself (e.g., "zerobranes"). Nonetheless, at least with the green patch, it is plausible to hold that if it is physical, it is extended in physical space. After all, thing is extended in a way. It exists as a blob with a certain size in the visual field. So if it is physical, its size is presumably its physical size, meaning it would occupy some region of spacetime.

Frustratingly, however, the green thing is apparently not something you can locate in or outside the skull. For this reason, Dennett (1991; 2013) bites the bullet and denies the existence of this and any other quale. (See also Dennett 1988.) But as many have complained, Dennett's denial seems like sheer denial. The green thing may not exist in the same way that my green shirt exists. Still, there seems to be an obvious sense in which, among the various cognitive phenomena, there are after-images. Further, it would be difficult for Dennett to explain why (1) even *appears* true. After all, an appearance of green is precisely what's eliminated! (I doubt the point is original to me; one can imagine someone like Searle levying this criticism.)

Regardless, Dennett is hardly unique in his denial. Several philosophers seem to fall in line, one way or another. For instance, Sellars (1960; 1963) also denies (1), at least as formulated. He jettisons (1) in favor of an "adverbial" paraphrase, along the lines of:

(1*) I am experiencing greenly.

This is to be understood on the model of 'I am seeing clearly and distinctly' or the like. But unfortunately, the adverbial view of qualia faces grave objections (see especially Jackson 1977, Butcharov 1980). And since space is limited, I shall leave adverbialism aside.

Place and Smart also opt for a Dennettian denial, for they too reject (1) as formulated. They propose to replace it with a paraphrase along the lines of:

> (1**) I am in an experiential-state that is type-identical to the state that occurs when I see an actual green thing.

In this, there is no mention of the after-image as such, so no ontological commitment to the thing is incurred. But as one may guess, multiple realizability renders (1**) problematic. Even within a single individual, a quale can be realized by quite different types of neurological states. (In the case of "pain," compare the neurology of a stomach ache with that of stepping on a nail.[7])

How else might a physicalist resist the argument? It may be possible to raise doubts about (3).[8] Perhaps the after-image could be token-identical with some green-hued electrical impulse traveling along the neural net. But, assuming it is even possible for an electrical impulse to be green, it will likely not have the right shape. (My thanks to Chris Daly for this observation.) Further, it is unclear how one would accommodate qualia from other sense modalities. Must our neural impulses make little sounds and smells as well?

Perhaps a functionalist response to the argument is the best option. Assuming that crude machine functionalism is off the table, a teleofunctionalist might classify the after-image as a biologically abnormal token of a specific functional type. For instance, the after-image might be typed as a green color quale—and a green quale could be type-identified as a state that has the function of carrying information about green-instances in the environment. Then, the after-image could be identified as a token of that type, though it would be a malfunctioning one, a false positive. It would be a *mis*representation, per the teleofunctionalist's account.

I desperately wish that this view were adequate. But it all turns out to be a non-starter. Suppose with the teleofunctionalist that the green after-image is token-identical with a neurological state having such-and-such the function. Consider further that the green after-image is green. Hence—this is important—it follows *by the indiscernability of identicals* that your neurological state is green! But your neurological state is not green. So it is not identical with the green after-image. Maddeningly, this argument seems decisive.

Lycan (1987; 1996) is also a telofunctionalist, though he offers a different reply to the Place-Smart argument. He proposes to classify the green after-image as an "intentional inexistent," in the sense of Chisholm (1957). In so classifying it, Lycan identifies the green thing as the

[7] Accordingly, I am agnostic about Kim's (1992) species-specific reductionism. Within higher organisms, "pain" is too coarse-grained to expect a non-wild disjunctive reduction, as the above illustrates. But if the types are too fine-grained, then laws about the types are less commonly instantiated and are thus less applicable for explanation or prediction. And it is not apriori obvious whether there is a level of grain that is "just right"

[8] See, e.g., Sellars (1962, p. 37). Other philosophers claim that *veridical* qualia are in the world (see, e.g., Tye 2009). However, the case of the after-image is why the Place-Smart argument is so forceful—there is no worldly green-instance in this case.

intentional object of a mental representation, albeit an object that does not actually exist. The green thing is thus akin to Pegasus or the fountain of youth. In accord with (2) and (3), none of these objects are located in space—yet they are still objects of some thoughts.[9]

Note well, to be an object of an intentional state is not necessarily to be an object of a *conscious* intentional state. For Lycan, a state is conscious only if there is a higher-order representation of the state, and not every representation of a green color-patch is itself represented. This is as it should be. Presumably, a green quale partly explains why Keona drove away from the stoplight when she did—even though she was slightly distracted and did not consciously register its occurrence. (But rest assured; Keona is generally an excellent driver.) Yet it would remain a case where the changing stoplight caused some non-conscious representation of a green color-patch (though here the intentional object is also an object in the world).[10]

Lycan admits, however, that the mystery of the after-image is “solved” only by assimilating it to another mystery, that of intentional inexistents. Still, he thinks that this is progress—one mystery is better than two. Nonetheless, it means that the Place-Smart argument remains unanswered: All premises are left standing, and ‘intentional inexistent’ is just a new label for the thing that exists nowhere in space.

### *3. Merely Intentional Objects*

However, Lycan has a further problem. His view implies that the green color-patch does not exist—it is an intentional inexistent. So why is this any different from a Dennettian denial? Why isn’t this just eliminativism about qualia?

Well, Lycan isn’t an eliminativist. But then, in what sense is the green color-patch *nonexistent*? Forget whether the thing is consistent with physicalism. The problem now is that the view seems internally inconsistent—the color-patch both exists and does not. Lycan seems to be with Meinong in saying: There exist things that do not exist.[11]

[9] Is the green after-image really the *referent* of a mental representation? Since it is an “image,” it may be more tempting to see it as the mental representation itself, one that refers to a non-located green thing. There may well be such a distinction between the image and the thing. Yet which of these would be the *quale*? As far as I am concerned, you can regiment the terminology according to taste. For my real concern is with the non-located green thing, by whichever name you call it. I will thus continue to gloss any distinction between the after-image and the color-patch/green blob/quale. (Relatedly, the green object might be the *content* of a mental representation rather than its referent. But here too, it does not seem to make a difference, re: the Place-Smart argument. Thanks to Kelly Trogdon here for discussion.)

[10] A further, useful distinction in Lycan is between the quale and the what-it’s-like feature of the state. The latter is a property of the experiential state as a whole, whereas the quale is just a proper part of the state. (Hence, as is suitable, it makes sense to talk of what it’s like to experience a green quale.)

Question: Is a non-conscious color-patch truly a *quale*? Since ‘quale’ is a technical term, the matter can be somewhat stipulative. But since the color-patch seems non-locatable regardless, my use of ‘quale’ is indifferent to the conscious/non-conscious distinction. (If pressed, we could coin a new term for non-conscious vs. conscious color-patches. But let’s not.)

[11] There is a particular irony here, since Lycan (1979) notoriously declares unregenerate Meinongianism to be “literal gibberish.” (There too, the issue was the apparent internal inconsistency.) But I hardly mean to be flippant toward Lycan. The reader will easily verify that his views have been highly influential upon the present work.

Perhaps we can avoid putting things in such dire terms. Instead of describing qualia as nonexistent or nonphysical, we might start by describing them as mind-*de*pendent objects, in contrast to mind-independent objects like rocks and houses. In the present context, to say that an object is mind-dependent is to say that it is a *merely intentional object* or "MIO." It is an object of thought, albeit merely an object of thought. It is not an object that can exist outside the mind. Fictional objects, hallucinated objects, etc., would be further examples of MIOs. Note, however, that artifacts like tables, the U.S. Constitution, and so forth, are not "mind-dependent" in this sense. After all, tables are not merely the object of intentional states; they also exist in the external world. (Perhaps the table is not really a *table* unless there are minds to assign it a certain dining-function. Still, the thing that is an actual table also exists in worlds without minds, even though it might not exist in those worlds *as* a table.)[12]

Calling such objects merely intentional is, I think, informative in certain ways. But ultimately these descriptions do not remove the contradiction; they only hide it. For when we described MIOs, we described them as objects of thought which do not exist "outside" the mind. Yet this already implies: There exist objects of thought which do not really exist!

William James once said that when faced with self-contradiction, make a distinction. I shall follow that advice here. After all, from one angle, it is virtually a Moorean fact that there are fictional characters such as Hamlet, Sherlock Holmes, and Sponge Bob. *And if we find the Dennettian denial intolerable, we are similarly committed to the after-image.* However, in light of the Place-Smart argument, it seems we must say that after-images, like fictional characters, do not really exist, i.e., they do not exist in physical space. But here is where the Jamesean tactic enters: We distinguish between existing as a MIO and existing in physical space. We are thus attempting consistency in the statement that MIOs exist as MIOs, even though they do not exist as spatial occupants.

The two senses of 'exist' clearly warrants an extended defense, and for this, I must refer the reader elsewhere (Parent ms.). Yet let me note that the distinction is not just philosophers' jargon. It is not esoteric philosophy-talk to say that some things are the stuff of myth, and some things are not. Apparently, such a distinction is already present in ordinary English.

If there is a distinction in senses of 'some', 'exist', and the like,[13] then quantification in the Place-Smart argument is potentially equivocal. (1) is true if 'there is' is a quantifier ranging over MIOs. But it is false if it is a quantifier ranging just over things located in physical space. Similarly, (5) is true if 'there is' ranges over MIOs, but not if it ranges only over objects in space. The consequence is that physicalism remains viable if the range of the quantifier includes only mind-independent objects. But here is the rub. If we are dead-set against Dennett's rejection, then we remain ontologically committed to a green thing that exists nowhere in physical space. So even if the Place-Smart argument has two interpretations, there remains one univocal interpretation which *still* seems to force us into dualism.

[12] Are minds themselves mind-dependent in the current sense? I am a quietist on this matter; see below for more.

[13] I assume here for simplicity's sake that 'exist' occurs in logical form as a quantifier rather than a predicate. Even so, I may ultimately sympathize more with the predicate view in, e.g., Azzouni (2004).

In fact, the distinction in senses of 'exist' may just make things worse for the physicalist. For there is now a way of arguing that it is *logically* impossible for the after-image G to be token-identical with something physical P. Suppose otherwise for *reductio.* Then, since P is not a MIO*, it follows by the indiscernability of identicals that G is not a MIO.* But according to the present view, G is a MIO—it exists as an intentional object, yet it does not exist beyond that. By the simplest sort of *reductio*, then, G cannot be token-identical to any physical stuff!

Before pursuing the dualism issue further, note that (1) was also said to be *false* if the quantifier ranges just over physically located objects. So on the alternate reading, the Place-Smart argument apparently forces us into *eliminativism* about qualia. However, eliminativism of one sort should come as no surprise. Suppose with David Lewis (1986) that our quantifiers can be more or less restricted in different contexts. Then, unremarkably, there will be contexts where green after-images are not in the range of the quantifier. For instance, in some contexts, the quantifier ranges over only things in my fridge. ("There's no beer [in my fridge]!") So thanks to quantifier restriction, it is only natural that context sometimes renders eliminativism a matter of course. ("There's no green thing [in my fridge]!")

One may protest that this is a "superficial" sort of eliminativism—for the unrestricted quantifier still has (or ought to have) qualia in its range. And the unrestricted quantifier indicates what we are *really* ontologically committed to. For present purposes, that can be granted. The point is just that the Place-Smart argument at best supports this superficial kind of eliminativism, the kind that results merely from quantifier restriction. For we are agreed that there is a less restricted quantifier where (1) is true.

I thus propose to leave aside the pro-eliminativist argument. What is more troublesome are the pro-dualist arguments, where our quantifier ranges over a physically unlocatable after-image, one that logically cannot be token-identical to anything physical. What is a physicalist to do?

## *4. Toward Theory Dualism*

As advertised, my aim is to interpret the situation not as demonstrating *metaphysical* dualism, but as rather as reflecting a dualism of *theory*. The idea is that there are principled reasons why we cannot combine a theory of MIOs with a theory of non-MIOs into a consistent whole. If so, then the mistake in the Place-Smart argument is to assume we must put have green thing be part of the *same* theory that includes physical things. But in fact, (1) on the relevant reading should be quartered off from physics into a different theory—and never shall the two meet. Though again, what explains this division of theory need not be a metaphysical division in reality. Instead, I suggest that the division owes to an in-principle limit on representation, one explainable by Russell's vicious circle principle.

If physics and the phenomenal are separated into different theories, then the contradictions cannot be derived from any one theory. This may not to cure our cognitive dissonance, however, since both theories are ones we accept. And though the conflicting statements are quartered off into different sets, their union still engenders contradiction. The division into different theories thus seems like a superficial fix.

There is a question begged in this, however. The objection presumes that a single theory of reality is the best way to think of our commitments. But theory dualism is opposed to exactly that idea. The theory dualist holds that the best way to regiment commitments is to *divide them*—we must separate statements about physical objects, from statements about MIOs such as after-images. The "separation" means that, in practice, it is as if the objects of one theory do not exist when operating from the other theory. Thus if we are restricted to the domain of physical objects, it is as if eliminativism is true. But analogously, while locked into the Husserlian *epoché*, physical objects are not to be found. Regardless, if we insist that after-images are a striking, non-trivial fact about perceptual cognition, then the separation from physics need not compel total rejection. Even if one theory is suspended when the other is operative, we might insist on the importance of both theories, so that neither is ultimately eliminated.[14]

This maneuvering is not artificial. In a given context of inquiry, our assumptions are often a restricted set of what we might otherwise assume. As a descriptive point, this is an undeniable fact about scientific practice. But as a prescriptive matter as well, differing assumptions are often rationally required by differing inquiries. One should *not* be ontologically committed to electrons, when the aim is to conduct an unbiased *test* regarding their existence.[15] Yet if electrons are part of our best current science, then we should refer freely to them when (e.g.) attempting to explain magnetism. Depending on the results of the former inquiry, the latter might take different directions. But it is possible for both inquiries to be worth funding, even though they are inconsistent on the commitment to electrons.[16]

Theory dualism is also not an artificial maneuver insofar as it is naturally emerges from the "science language game." Borrowing a metaphor from other work (Parent ms.), one might compare scientific activity to the action of a coin-counting machine. The machine first separates out the slugs and other noise, and then categorizes what remains as quarters, dimes, nickels, etc. In a similar way, normally science first removes from our impressions what is fictitious, hallucinated, or otherwise illusory. The remainder is then classified into natural kinds.

If the coin-counting goes properly, the end result will be as if there were never any slugs in the initial input whatsoever. Similarly, in properly functioning science, the final product will be as if MIOs never existed. In both cases, there are sound reasons for designing the process that way. Yet in neither case should we be fooled into thinking that there are no "counterfeit" objects to begin with. Indeed, part of the point of the coin-counting machine is to weed out the slugs and other noise. In the same way, part of the point of scientific activity to eliminate what is fictitious, hallucinated, or otherwise illusory. Once that is appreciated, it should be no mystery why after-images and the like are absent from the domain of physics. The scientific process is *by design* one that eliminates such things, prior to any further descriptive work.

[14] Here and elsewhere in the paper, there are connections with a view called "mental fictionalism" (cf. Parent 2013). Originally, my plan was to discuss these connections, but I have already gone over the word limit. Hopefully, I will be able to discuss these connections in future work.

[15] This is so, if ontological commitment means that the existence of electrons would be *assumed* in the testing-context. However, one's present ontological commitment to *x* does not mean one cannot gather fresh evidence on whether *x* exists. The point is just that the commitment would have to be suspended if one wants an unbiased test. (Thanks to Chris Daly for raising this issue.)

[16] There is more to debate, re: contextualism about ontological commitment. I discuss such a view further in §4 of Parent (2008). But a proper discussion requires its own paper, one which I hope to write in the near future.

Consequently, it is only natural that MIOs do not "exist" in the standard, actualist sense—where the quantifier ranges over the proper domain of natural science. *Of course* there are no fictions, illusions, etc., as far as science is normally concerned. Nevertheless, it is partly because there are MIOs that disciplined scientific inquiry is needed. It may be thought "anti-scientific" to admit their existence. But quite the contrary, one cannot properly appreciate all that science does for us, without recognizing how it protects us from phantasms of our own making.

Still, there is something quite foreign about theory dualism. Assuming there are truths about qualia, the default is to assume that they could be integrated into a single, consistent theory with all the truths about the rest of reality. It turns out, however, that this is mistaken. As concerns qualia, we are dealing with a case where there are in-principle reasons why inconsistency with other scientific domains is inevitable. If so, then on pain of contradiction, the best we can do is separate truths about qualia into a different theory than physics, and maintain silence on how truths of the two theories are related. A consequence is that *no single picture of the world* is provided by theory dualism. The theory dualist acknowledges the importance of both physics and phenomenology, yet recognizes that "bridge laws" joining the two theories would only breed contradiction. The theory dualist is thus resigned to an incomplete understanding of the world. She is a *quietist* how qualia are related to the domain of physics—and this applies to (2) and (3) in the Place-Smart argument. She thinks there are in principle obstacles to making sense of such claims.

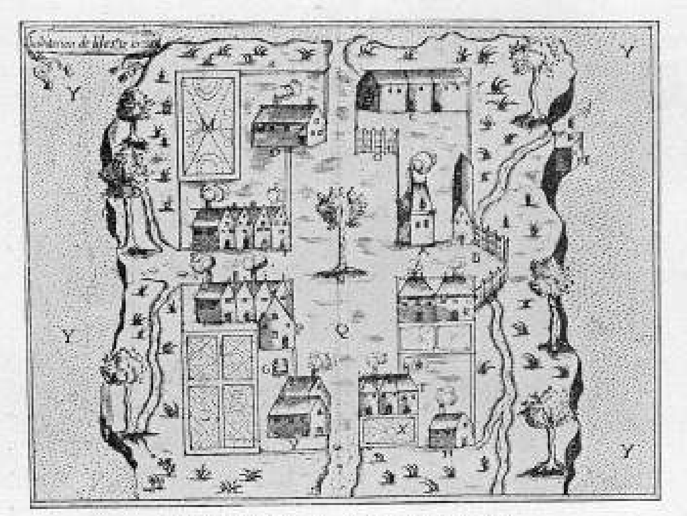
Now that the idea is on the table, theory dualism can seem *ad hoc.* At worst, it would allow one to assert any contradiction with impunity, as long as the contrary statements are "quartered off" into different theories. Yet this is why the theory dualist has to be clear on why this is a special case. It is agreed that not every such separation would be well-motivated. But with qualia, there is independent reason to think that paradoxes will arise, *even in a physicalist universe,* having to do with in-principle limits on representation.

### *5. The Map of St. Croix Island*

Since the argument will be somewhat technical and abstract, this section and the next are designed to foster a more intuitive grip on the situation. To this end, I shall develop an analogy, meant to illustrate the ways in which a physicalist might try to represent the very representations she uses.

Suppose that we have a grayscale map of St. Croix Island (a.k.a. Dochet Island). And imagine that we are standing on the island, looking at the map. Now the map shows the location of various points of interest, and is informative of some of their properties. E.g., the map tells us that the former dwelling of Samuel de Champlain is at such-and-such location. Suppose further that we pinpoint on the map where we are. But then I turn to you and say: Where on the map does *this very map* appear? (The map presumably will have no such indication of itself.)

My question is an odd one, but suppose you indulge me. You draw a small square at our location on the map, stipulating that the square represents the map on the map. However, suppose I insist our map should be represented in complete detail on the map —the square ought to be modified to represent *every* fact about the map itself. *Could this be done?* There is reason to think it could not—not even by God. The quick argument is this: Every modification to the square would itself be a new, additional property of the map that also needs representing. In adding a symbol, we may thereby capture one feature of the map, but at the cost of adding a new feature to be represented, viz., the symbol just added. If we are stuck in this trade-off, then no matter how many symbols to the square are added, there would always be some further fact about the map to represent. We would be caught up in a regress, and we would never achieve a fully self-representing map.



BUILDINGS ON STE. CROIX ISLAND, 1613
(From Champlain's diagram)

A map of Champlain's settlement on St. Croix Island.

There may be ways out of the regress, but first, why would I want a fully self-representing map at all? Well, I might start by observing that the map is one thing, and the island is another. And I do not want my map to suggest a kind of "map-island dualism." But map-island dualism might be implied if the map is omitted from the map, as if the map exists without any specific location on the island. Thus, you graciously draw the square on the map. Yet then, I notice that many features of the map remain unrepresented. Hence, although your drawing rules out map-island *substance* dualism, map-island *property* dualism still seems live. If the map's own property-instances are absent from the map, then that might suggest, as before, that some of these instances have no location on the island.

Thus, to represent the property-instances, suppose you start by drawing a small house inside the square, to represent that the map has a symbol for Champlain's house. You have then captured one property of the map—that it has a figure of a house—but you have also bestowed a new property on the map. The map now has a second house-figure inside the square. Here arises a choice. First, one could represent the second figure by drawing a third—but that leads to a house-figure regress. The other option is to leave the second house-figure being unrepresented. That may leave open the possibility of "nested-map-island dualism," but perhaps that is a case of "don't care." As long as the map avoids "first-order" map-island dualism, we might be reassured that in principle, it could rule out second-order, third-order, etc., dualisms as well.

But to the contrary, if we rest content with an unrepresented nested figure, first-order dualism may remain live. Suppose the unnested house-figure is drawn in black ink. Still, it may be understood that the figure does not represent the house as black—we can stipulate as part of its representational powers that the black figure represents (say) a red house. However, the color discrepancy between the symbol and the house is one I may find important. So suppose I want to represent on our map that the house has a different color than the figure. How would we do that?

Presumably, we would draw the nested house-figure in a way that represents the first figure as having a different color. But suppose we only have black ink with which to draw. Then, if the black ink of the first figure represents red, the black ink of the nested figure may *mis*represent the non-nested figure as red. Accordingly, the black property-instance of the non-nested figure remains unlocated on the island—meaning that map-island property-dualism remains in the running. To preclude this, I might add a note on the map that the black ink represents redness in one case, and blackness in the other. Yet then, my note introduces further symbols on the map. These too might be represented by further symbols, but now I have restarted the regress.

So the attempt to represent the Champlain house as a different color from the house-symbol is one way the regress can continue. However, a way out may lie in the use of *self-referential* symbols. The hunch is that thanks to self-reference, perhaps the regress could terminate in referential circles, rather than continue ad infinitum. But how exactly would this play out?

Imagine for instance that the second house-figure (inside the square) is interpreted as a symbol for itself. The problem, however, is that it then no longer represents the first house-figure (the one outside the square). Or at least, it no longer represents the first in a *consistent* way. It might be allowed to stand as a representation that inconsistently represents both figures. But if we want the map to represent all of its own properties at once, then "the" object of the nested figure will be represented as something both nested and non-nested.

A different self-reference strategy is to code the entire map by a single numeral, say '42', writ on the spot corresponding to the location of the map. Then, '42' can be seen as representing everything about the map, including the fact that '42' is written at that spot to code all the facts about the map. (One posits some function that maps '42' onto the totality of facts about the map.) The problem, however, is that a user is unable to recover the facts encoded by '42', unless she independently knows of those facts. For instance, one could not learn from the numeral '42' alone that the map has a house-figure representing the Champlain house. That is so, unless one knows already that the map has such a figure, by independently inspecting the map. So it seems one cannot *learn for the first time* what the map looks like via the numeral alone. The code is thus of limited use as a representation. It is like visiting St. Croix Island for the first time where, instead of being handed a map, the park ranger at the entrance just tells you "42."

Here is a different self-referential tactic. Perhaps we can forgo all modifications whatsoever to the map, and simply stipulate: Let the map *itself* represent the map. Thus far, the map's location on the island goes unrepresented, yet we could add the drawn square to fix that. But then we run into a problem akin to that of the self-referring house-symbol. In order to represent all the island-facts at once, the map must be used to represent the island *and* itself simultaneously. As before, such inconsistency leads to absurdity; in this case, the map can suggest that the map *is* the island.

So suppose instead we buy a second copy of the map, and use it as a representation of the first. (This is the analog of theory dualism, though it too has important limitations.) Now in this case, we can still draw a square on the first map to locate the first map itself. But as far as the first map is concerned, that square is a "black box;" it represents hardly any property-instances of the first map. Learning about these is possible, however, if we take the second map to be a representation of the first. The second map then acts like a detail or "zoomed in" representation of the thing plotted on the first map (viz., the first map itself).

The problem here, however, is that neither map represents the relationship between the maps (unless in a limited way, by some code). One *might* follow the mapmaker's practice of integrating two maps into a single layout—where an arrow is used to indicate that the second map details something that is plotted on the first map. Unexpectedly, however, this requires interpreting the second map in two different ways. On one hand, the second map details the first map (as plotted on the first map). But when the arrow is drawn between the maps, the second map is also being used to represent itself. After all, that is how the arrow can indicate that *the second map* details something plotted on the first map. Consequently, if one wants a representation of all the facts on the island at once, the second map requires an inconsistent interpretation. (In this instance, "the" object represented by the second map is identical to both the second map and the first map.)

So the lesson, apparently, is that we cannot obtain (at least in a learning-friendly and consistent way) a complete representation of the map on the map. In brief, either the symbols used would be partly self-referential or not. If they are non-self-referential, then each new symbol becomes another thing to represent, resulting in regress. Whereas, if a symbol is interpreted as (partly) self-referential, then it is either interpreted inconsistently, or the symbol will be of little use for learning. In any of these cases, however, there will be a limit on how things are represented.

## *6. Unpacking the Analogy*

This, I think, is analogous to the situation with qualia. The island as depicted on the map is analogous to the physicalist's model of the world. The problem presented by a green quale, then, is that it lacks a location in her model. This is like the map initially lacking a representation of its own house-figure. A further similarity is that the quale is used to represent something, viz., an instance of green, just like the house-figure is used to represent something on the island. Though a key difference is that the green-after-image is a *mis*representation, and the map was presumed to be accurate.

A misrepresenting quale was invoked in the Place-Smart argument, just to make clear that the green patch was not located outside the skull. A veridical green quale would have made that less clear—but ultimately the color of a veridical quale must also be distinguished from the worldly property-instance it represents. Briefly, if one is committed only to the worldly property-instance, then there is no longer anything *veridical.* (Admittedly, the distinction in the veridical case remains contentious, though I shall say more about it later.) In order to maximize the map-analogy, however, it will be best to consider a veridical green quale, say, the one that caused Keona to drive away from the stoplight.

Now Keona happens to be an astute physicalist philosopher. She thus aims to "locate" all of her qualia in her physicalist model of the world. In the case of the stoplight specifically, the green quale is a representation with something like the content "green over here," or "lo, a green property-instance." This "semanticizing" of qualia will be contentious in some circles—but unfortunately, I am unable to mount a proper defense here.[17] Yet I would note that "veridical quale" seems like a category mistake if her quale has no representational content. Still, if preferred, one can see my remarks as limited only to perceptual representations, whether or not these count as qualia. (Or rather, I will allow this, as long as the after-image would count as a perceptual (mis)representation of green.) But for continuity's sake, I will continue to speak of the "quale."

Thus, the quale can be seen as representing some feature/part of Keona's physicalist model of the world, or, the world according to Keona. This means that, even if the quale is not yet represen*ted* as something in her model, it is represen*ting* such a thing, namely, the greenness instantiated by the stoplight. But as a physicalist, Keona also wishes the quale to be represen*ted* as an object in the model. This, I suggest, is akin to representing features of the map using features of the map. And as in the map example, Keona's endeavor will encounter certain limits.

Here too, the options for representing the thing are: By non-self-referential means only, or by self-referential means at least in part. Suppose Keona attempts the latter. Then, she might try to represent the quale as part of her model by using the quale itself. Yet the quale then acquires an inconsistent interpretation. In particular, one can ask: Is the quale representing a property-instance outside her skull, or not? It is clear that she cannot answer "both." Otherwise, it follows that a single trope is and is not outside the skull.[18] Note that the situation is akin to the nested house-figure in the map, when it was hypothesized to represent itself *and* the non-nested figure. There too, the problem was the implication that a single thing had an inconsistent location: both outside and not outside the drawn square.

So it seems Keona cannot represent the quale in her model by using the quale itself. But what if she uses a symbol that is only partially self-referential? This would be like using '42' to encode all features of the map, including the fact that '42' appears on the map as a code for all those features. In Keona's case, she might also use the numeral '42' to code all the relevant facts—including the fact that all those facts are coded by '42'. Then, when she writes her complete "Book of the World," she could use '42' to register all those facts. The problem, however, is that as in the map-example, the relevant facts cannot be communicated to someone unlearned. For instance, when Jackson's (1982) Mary is locked in her black-and-white room, she would be unable to glean from Keona's "Book of the World" certain facts about the quale. To be sure, Mary may well understand that '42' is being used to represent "all the facts about Keona's green quale" (whatever those are). But the numeral alone would not communicate what those facts are, except in this generic, non-descript manner.

[17] Though see Lycan (2014), who argues that even smell-qualia are representational. But note well, I take "semanticized" qualia to be different from the raw input from (e.g.) the retina. Hence, I do not regard the above remarks as applicable to raw data, and admittedly, there are probably ontological difficulties about that as well. My thanks to Josh Entsminger for discussion here.

[18] I take such inconsistency to be an issue for "self-representational" approaches to qualia, as in Kriegel (2009). Nearby views may also be affected, e.g. Gertler's (2001) demonstrative view, Balog's (2012) quotational view, etc.

True enough, Keona is able to communicate more to Mary if her encoding is more sophisticated. ("The green quale is type-identical to the quale that results from looking at healthy grass in broad daylight)" But as is familiar from the Mary-literature, there seems to be a persistent failure to communicate *something* about the quale. It turns out, however, that this phenomenon is a special case of a broader one about the limits of representation. It is the same sort of thing that happens when the park ranger greets you at St. Croix island with the numeral '42'. Unless you know the facts encoded by the numeral independently, the numeral alone does not communicate those facts (except perhaps in a generic, non-descript way).

Yet unlike Keona's green quale, Mary can learn the facts about St. Croix by independent means, by exploring the island herself. But Mary herself cannot explore *Keona's* green quale. Keona's quale is not something one can visit. It is only an object of her "internal world," and qua internal, it is not something we can epistemically access. So there are really two points here:

(i) Some phenomenal facts cannot be learned via codes, and
(ii)Unlike facts about St. Croix Island, there is no other means for third-persons to learn of those facts.

Together, these imply that third-persons cannot learn all facts about the quale. Lycan (2003) once summarized this situation well in the following:

> phenomenal facts are ineffable. One knows them from the inside, under special introspective modes of presentation… and the representations in question are not synonymous with expressions of any public natural language, actual or possible. What is ineffable cannot be explained at all; at the very least, the introspective representations are not going to be deducible from microphysics or even from neuroscience or from any other body of public information expressed in public notation. Yet all of this is… compatible with materialism. (p. 139)

Lycan is right that ineffability and exclusive first-person access would not alone rule out physicalism. In this case at least, epistemology does not determine the metaphysics. It is also plausible that the elusive information consists in facts known under exclusively first-personal modes of presentation.[19] But I would also add something further to the story— that ineffability is really a special case of a more general limit on representation. In one sense, the phenomenal facts *are* "effable" if we just stipulate some code for them. But in that form, the facts are not represented in a learning-friendly way. And so in one sense, they cannot be "effed." Yet the misfortune is compounded, since unlike knowledge of St. Croix, there is no independent means to acquire it. For the facts are known only under a distinctly first-person mode of presentation, a mode of presentation available to the first-person alone.

Let me add that the communication of phenomenal properties is not the only issue here, perhaps not even the most interesting issue. In the map case, we were trying to avoid "map-island" dualism by performing a certain epistemic (or better, representational) feat. We were trying to represent features of the map on the map itself. In Keona's case, she is attempting an analogous

[19] This is part of Lycan's "phenomenal concept" strategy against Jackson, and I recognize that the view is not universally accepted. But the strategy will not be important to the remainder of the paper.

feat, due to an analogous ambition. She wants to avoid qualia-physics dualism, and her hopes lie in representing the green quale as a part of the physical world. This feat can be achieved by encoding, but in so doing, some facts are thereby unlearnable. Yet this point about learning is really more disconcerting than we have appreciated.

Matters are clearest if we momentarily switch back to the non-veridical after-image (though the point also applies to the veridical case). With the after-image, it is plain that no code will communicate a fact of which *everyone* is ignorant. Namely, it cannot communicate how something *not* in space can be made of ingredients that *are* in space. Even the first-person is ignorant of that. How is it possible for physical fields, forces, etc., to aggregate into a green thing with absolutely no location?[20] Or if we want to allow for the possibility of "neutral monism," how is it possible for one kind of stuff to compose things that are extended in physical space, *and* to "compose" other things that are not so extended?

If we code all the facts about the quale using '42', we ipso facto code the facts that these questions concern. But since we do not already know what those facts are, the code is no help in answering those questions. We would be trying to learn from a code a metaphysical relation that the code leaves implicit. More, neither first- nor third-persons have an independent way of "exploring" the relation between the unlocated and located things. If one tries to explore it using physics, one relatum is missing from the domain of inquiry, viz., the spatially absent green thing. Conversely, if one tries to explore via the Husserlian *epoché* or whatnot, the other half of the metaphysical relation is missing. And naturally, if we take the union of the domains, we have a domain where some green things have no physical extension, contra physicalism.

Like Mary, we want to know facts about a quale that meet conditions (i) and (ii), above. In both cases, the facts cannot be learned from a code, nor can they be learned by some independent means. Unlike Mary, we are not trying to learn "what the quale is like" or some such thing. We are instead trying to learn of the metaphysical relation between the quale and things in physical space. But the epistemic problem that afflicts Mary also afflicts our quest for metaphysical knowledge. That is how the epistemic issue comes to bear on metaphysics. We simply cannot answer, just on the basis of encoding or any other means, how exactly the after-image "fits in" to spatial order. (Even the metaphor of "fitting in" presupposes the quale is somehow in space!)

So again, a representation of a quale encounters limits if it is represented by itself, and also if it is represented by some partially self-referential code. But we have not yet considered representing the quale by purely non-self-referring means. Yet this is the case where the regress arises. What drives the regress, in one sense, is just the ambition to represent our entire representational edifice as part of the model represented. But this is not the best way of understanding what really vexes Keona. The problem lies not just in her being a completist about her collection of the facts. What is more disturbing is the apparent *metaphysical oddity* of her quale. Initially, the quale is not represented in her model, as if it did not occupy space. Thus, Keona attempts to incorporate it into the model. Suppose, for instance, she identifies the quale with a green spot on her retinal image. Here, she will run into a problem, akin to representing the house-symbol as black while representing the house itself as red.

[20] Given a measurement of momentum, even an electron is located in one sense—it is located in *multiple* places, as per some probability distribution. But the after-image fails to be located even in that sense.

Recall that the black ink of the unembedded figure represented the redness of Champlain's house. But for that reason, it was problematic to use black ink for the embedded house-figure, since that can misrepresent the color of the unembedded figure. This prompted us to suggest that the color of the unembedded figure might represent itself as well as the color of the house. But since the two objects are different colors, we landed in contradiction.

The inconsistency is also present if the black property-*instance* of the unnested figure were to represent itself and the red property-*instance* of the house. This switch to color-instances makes the analogy with Keona tighter. For in Keona's case, she is representing an instance of green on the retina as identical with the quale, so that in her model, one instance is numerically the same as the other. But in the end, the two "greens" cannot be identified. For it is possible for one green-instance to occur in the absence of the other. Foveating a green stoplight can fail to cause a green quale, e.g., in a person with red-green colorblindness. Even so, the person's retinal image will have a patch of green. So the greenness of the quale is not the greenness of the image.

Hence, some third location for the instance of phenomenal green needs to be found, and the regress begins. Actually, it is hard to see how to continue regress in a plausible way, once the retinal spot is discounted. But the overarching point is that we are encountering a limit on representation here, which bears some analogy to the problem with the map. In both cases, we want to represent a representation as, itself, part of the domain represented. But the thing that represents the representation is, without further comment, unlocated. And if we represent it as having the color of some located thing, we seem to misrepresent the facts.

Earlier, we noted a theory dualist tactic of using a second map to represent the first. What is the analog in Keona's case? The fact is that she always had two conflicting theories; the problem was to unify them into a consistent whole. So the theory dualist just advises surrendering that ambition, thus allowing the theories to stand as if they were representing two different realities.

But the "as if" is important, for it reminds us that metaphysical dualism does not strictly follow from theory dualism. The analogy is that, even if both copies of the map are not *represented* as located on the island, both copies of the map still *are* on the island. The metaphysical thesis does not follow in either case, just because our representation of the metaphysical facts is incomplete. This also provides a hint as to why 'There exist things that do not exist [in space]' is true on one interpretation. Keona's color-representation, used to represent part of the world, is not itself represented as part of the world.[21] The analog would be "There are symbols on the map whose location is not plotted on the map."

[21] To be clear, 'There is a quale that does not exist [in space]' does not have the same truth-condition as 'There is a color-representation which is not itself represented'. (Relative to a model, the former contradicts physicalism, while the latter does not.) Still, the truth of the latter may explain the truth of the former, assuming these truths nomologically covary. (Cf. Azzouni's 2010 notion of a "truth-inducer.")

## *7. The Argument for Theory Dualism*

### *7.1 The Heterological Paradox*

A green quale is like the house-figure in that both are representations we want represented, as part of the represented domain. If one can accomplish this, it constitutes a non-dualist picture of things. If the map is adequately represented by the map itself, the map thereby represents the island in a way that precludes map-island dualism. Similarly, if the quale is adequately represented by the Keona's theory, she achieves a non-dualist representation of the world. However, we have considered various strategies for constructing a thoroughly self-representing map/theory, and have encountered limits in each case.

This is no accident. Circularity, incompleteness, self-reference, and regress often are symptoms that a logical paradox is in the area—and the suspicion is correct in this case. The paradox here is an instance of the heterological paradox, and the derivation of the paradox is a clean, rigorous way to show that it is not possible to have a non-dualist theory of a domain, if the domain includes the representations used in your theory. Yet the derivation is not the best way of explaining *why* there is a paradox at all. The analogy with the map is probably more helpful on this score. The analogy clarifies just how problematic it is to plot all property-instances of the map on the map. For at every turn, we encountered regress, circularity, incompleteness, or inconsistency. In a way, that is more explanatory than isolating the oddball paradoxical sentence in the middle of it all. Though again, the paradox is more effective at showing that the problem is real. (Deriving the paradox might be compared to rupturing a dam by putting pressure on the weakest point. But if one asks why the dam was in peril at all, it can be more explanatory to point out its general deteriorating state.)

However, besides providing a demonstrative argument, the derivation of the paradox is explanatory in some ways. We shall appreciate that the paradox can be modified to show that the *location* of some representations cannot be represented. I take this to be partly explanatory of why, in particular, the location of the after-image is problematic. Similarly, there is a version of the paradox suggesting that some facts about *color* cannot be represented. There is also some explanatory value in this, regarding our confusion abut whether the greenness of a veridical quale is the greenness of a worldly object (another reoccurring issue in the previous sections).

In the general case, the paradox is derived from assuming that one can represent every fact about the map on the map, or that one can represent every fact about the physicalist's representations, using such representations. In particular, consider that the unembedded house-symbol has the property of being *non-self-denoting* or *heterological.* The figure does not denote itself; instead, it denotes the Champlain house. Moreover, lots of other symbols on the map are heterological. In fact, we saw that plausibly any self-denoting symbol ultimately would not serve our goal, so it may turn out that each symbol is non-self-denoting. But let us not prejudge that. The point is just that some symbols on the map are heterological.

Hence, if the aim is to represent every fact about the map on the map, we would need some predicate-symbol 'H($x$)' that is satisfied by exactly the heterological symbols, i.e., a symbol that denotes the symbols that are non-self-denoting.[22] Similarly, for the physicalist to represent every feature of her representations, she would require such a predicate as well. But the paradox arises when we ask: Is the predicate 'H($x$)' heterological? One can readily appreciate that the answer is "yes" iff the answer is "no."[23] So if consistency is mandated, the physicalist cannot include 'H($x$)' among her symbols. But again, 'H($x$)' is needed to represent all facts about her representations. Hence, some facts about her representations must be left unrepresented.

Thus far, this is not an argument for theory dualism, as much as an argument against theory monism. The argument does not automatically force us to be theory dualists. For we need not start talking about the heterological symbols of one theory, using a different theory—we can instead just maintain silence. However, I presume it is advantageous to say more if we can. (Even if some matters must be passed over in silence, one still hopes to say as much as possible.) In this case, "saying more" would mean rejecting the predicate 'H($x$)' as defined, yet still constructing a metatheory where some predicate 'G($x$)' is defined just on the *first-order* heterological predicates. (This is parallel to the type-theoretic solution to liar-like paradoxes and related phenomena.[24])

*7.2 Heterologicality and Color*

The paradox depends on the following definition of 'heterological':

> (D1) A predicate 'F($x$)' is heterological iff ~F('F($x$)').

As mentioned, there is another paradox concerning color that seems partly explanatory of our confusions about phenomenal vs. physical color. The paradox here is roughly a case of the heterological paradox, for it depends on the definition of a predicate that strongly parallels (D1):

> (D2) A color denoted by 'C' is a "non-self-exhibiting color" in a context iff, in the context, there is a token $y$ of 'C' such that ~C($y$).

Intuitively, a color denoted by 'C' counts as non-self-exhibiting relative to a context iff the context features a token $y$ of 'C' which, itself, lacks the color—i.e., $y$ fails to satisfy 'C'.[25] To illustrate, consider that the token of 'red' ending this sentence is not red. So red is a non-self-exhibiting color in the present context—some token of 'red' is a counter-example to the claim that all tokens in this context are red. Whereas, black in this context is a self-exhibiting color. All the tokens of 'black' are black, thanks to the use of black ink (or black pixels, for online readers).

[22] Note well: On my usage "$x$ denotes $y$" is true iff $y$ is a *member of* the extension of $x$. This is in contrast to some authors, who identify $y$ as either the extension itself, or as the plurality of members of the extension.

[23] If 'H($x$)' denotes itself, then by definition, it is not heterological. But if it is "not heterological," then the predicate self-denotes, meaning it *is* heterological. So the predicate 'heterological' is heterological iff it is not; contradiction.

[24] It is not entirely analogous, however, since I am thinking of the models for the two theories as completely non-overlapping. One model has an ontology of mind-dependent objects, and the other does not. Whereas in type-theory, the domain for an $n$-order theory is often a proper subset of the domain for the $n+1$-order theory. (Only the latter contains the $n$-order semantic terms, if any.) Unfortunately, I lack the space to discuss these differences here.

[25] N.B., if no token of 'C' occurs in the context, then vacuously the color fails to be "non-self-exhibiting."

Weirdly, however, we can show that black in our context is also a *non*-self-exhibiting color. Consider first that, if 'non-self-exhibiting color' = 'C' in (D2), we obtain that the term denotes a non-self-exhibiting color in this context iff some contextually-present token of 'non-self-exhibiting color' *fails* to have a non-self-exhibiting color—i.e., a token has a self-exhibiting color. But by the meaning of 'self-exhibiting', that implies that a token of 'non-self-exhibiting color' has a color denoted in this context by 'non-self-exhibiting color'. Yet since all such tokens are black, this means 'non-self-exhibiting color' denotes black in this context. So, black is a non-self-exhibiting color in the present context; QED.

This illustrates the problem on linguistic representations, but it is equally pertinent to mental representations. In particular, the contradiction arises regarding Keona's quale. Suppose that a token of GREEN is her mental representation of the stoplight's color. In Keona's context, then, is there a token of GREEN that fails to be green? Or is green a self-exhibiting color in this context? (Assume there is only one token of GREEN in the context, for simplicity's sake.)

Since Keona is a physicalist, she will judge the thought: GREEN IS A NON-SELF-EXHIBITING COLOR IN THE PRESENT CONTEXT.[26] This is because she thinks her quale is token-identical to some neural-functional state, and the neural-functional state is not green. However, the key idea is that a paradox arises *regardless* of whether Keona takes this physicalist stance. We can show that such a judgment is true iff it is false, because the concept of a "non-self-exhibiting color" is inconsistent. This is crucial, since it means that *independently of physicalism*, we inevitably land in contradiction on issues about color. That explains, I think, why such issues exist, why they should not tell against physicalism, and why we should ultimately be quietists on these matters.

To show that "non-self-exhibiting color" is inconsistent, consider that the following thought is true, assuming (D2). (Here, chevrons are the mental analogue of single-quotes; they are used to denote the mental representation named within the chevrons):

> (D2*) A COLOR DENOTED BY «NON-SELF-EXHIBITING COLOR» IS NON-SELF-EXHIBITING IN A CONTEXT IFF, IN THE CONTEXT, SOME TOKEN OF «NON-SELF-EXHIBITING COLOR» HAS A [NON-NON-]SELF-EXHIBITING COLOR.[27]

The truth of the thought follows from (D2), assuming that the thought expressed by (D2) is true given that (D2) is—and given that the thought expressed by (D2) entails (D2*). (Notice there is no presumption that (D2*) is an occurrent thought of Keona's. But it is at least an occurrent thought of mine, and the present claim is just that its truth follows from (D2).)

Nevertheless, in Keona's context, there is a token of NON-SELF-EXHIBITING COLOR, given her physicalist judgment about green. (We are not assuming the *truth* of the judgment, but we are assuming that such a judgment *exists*.) Since this is the only such token in Keona's context, it

[26] As usual in the literature, I use small caps to denote mental representations with the same content as the small-capped expressions. This need not be an endorsement of Fodor's (1975; 2009) "language of thought" hypothesis. It requires assuming neither universalism about concepts, nor computationalism about cognition, nor nativism about primitive concepts. The notational convention assumes only that there are concepts (whatever those are), and that these compose to form thoughts (whatever *those* are).

[27] Observe that (D2*) must be a *thought*, since per (D1), the very same symbol mentioned on the LHS must be *used* on the RHS.

thus follows from (D2*) (and from the fact that the left-hand side of (D2*) is trivial) that Keona's token of NON-SELF-EXHIBITING COLOR has a self-exhibiting color.

But, by the meaning of 'self-exhibiting color', it follows that her token of NON-SELF-EXHIBITING COLOR has a color that the token denotes. Since the token (trivially) denotes non-self-exhibiting colors, it follows that Keona's token has a non-self-exhibiting color. But that contradicts the conclusion of the previous paragraph.

So her token of NON-SELF-EXHIBITING COLOR has a color that is both self-exhibiting and not. In brief, what this illustrates is that representing the colors of one's color-denoting tokens is an impossible task. It may seem like a legitimate endeavor initially, just like it initially seems possible to to travel backwards in time and change the past, or to say whether any given term is heterological. But in all such cases, the prospect breeds paradox. In the present case, a token of NON-SELF-EXHIBITING COLOR ends up having a contradictory color. Thus, if Keona uses such a token to judge the color of a color-denoting token, any apparent success rests on her failing to recognize that *that very token—one that makes her judgment possible—could not have any possible color.* Or to be more precise, the color of the token cannot be consistently represented in her model (and that is so, regardless of whether it is a physicalist model or not).

Assuming inconsistency is not to be tolerated, a map or theory that aims to represent every color-fact about its domain cannot use the notion of a "non-self-exhibiting color." But then, some color-facts will go unrepresented. More broadly, due to the heterologico-color paradox, we cannot have a complete and consistent theory about the color of our color-representations. *In particular, we cannot have a consistent theory of what color some of our color-representations have.* Except for the last sentence of the next subsection, this is the most important idea in this paper.

### *7.2 Heterologicality and Location*

The other version on the heterological paradox seems to explain partially the conundrum of the Place-Smart argument. For the paradox here suggests that the *location* of some representations must be unrepresented. That, it seems, would explain how we run into a problem about locating a quale, *even if the universe is entirely physical*.

The location version of the paradox exploits the following notion of a "non-self-plotting" token:

> (D3) A location denoted by 'L' is a "non-self-plotting location" in a context iff, in the context, there is a token $y$ of 'L' such that ~L($y$).

Intuitively, a location denoted by 'L' counts as a non-self-plotting location in a context iff the context features a token $y$ of 'L' such that $y$ is not located at L—i.e., $y$ fails to satisfy 'L'. For example, consider that line 1 on this page is non-self-plotting in the present context: There is a token of 'line 1 on this page' in the context which is not located on line 1 on this page. Whereas, line 5 on this page is self-plotting in the present context. All tokens of 'line 5 on this page' in the present context occur on that line, i.e., the line above this one.

Suppose that ‘GREEN-LOCATION’ names Keona’s mental representation for the location of the stoplight’s color. In Keona’s context, then, is there a token of GREEN-LOCATION that fails to be at L? Or is L a self-plotting location in this context? (Suppose there is only one token of GREEN-LOCATION in the context, for simplicity’s sake.)

Keona is not some sort of Hegelian idealist; so she thinks the location of the quale is not the location of the worldly green-instance. Hence, she will judge the thought:

> GREEN-LOCATION IS A NON-SELF-PLOTTING LOCATION IN THE PRESENT CONTEXT.

Yet the judgment will be true iff it is false, simply because the very idea of a “non-self-plotting location” contains a contradiction. To show this, first suppose that:

> ***l*** = the location of Keona’s token of NON-SELF-PLOTTING LOCATION in the above judgment-token.

The contradiction is brought out when asking: Is ***l*** a non-self-plotting location? *The answer is ‘yes’ and ‘no’.*

Proof: Observe that (D3) implies the truth of the following thought (when ‘L’ is replaced by ‘NON-SELF-PLOTTING LOCATION’, and chevrons are the mental analogue of single-quotes):

> (D3*) IN A GIVEN CONTEXT, A LOCATION DENOTED BY «NON-SELF-PLOTTING LOCATION» IS A NON-SELF-PLOTTING LOCATION IFF SOME TOKEN OF «NON-SELF-PLOTTING LOCATION» IS SUCH THAT THE TOKEN HAS A [NON-NON-]SELF-PLOTTING LOCATION.

Note that the left-hand side of (D3*) is trivial. Hence, the right-hand side is actually true.

Furthermore, *ex hypothesi* there is only one token of NON-SELF-PLOTTING LOCATION in Keona’s context. So, that token must be what satisfies the right-hand side of (D3*) in her context.

That means: Keona’s token actually has a self-plotting location, i.e., ***l*** is self-plotting.
*But we can also show the opposite.*

If the token has a self-plotting location (per above), then by the meaning of ‘self-plotting location’, the token occupies a location that the token denotes. So, Keona’s token denotes ***l***.

Yet trivially, her token of NON-SELF-PLOTTING LOCATION denotes non-self-plotting locations. Hence, since the token denotes ***l***, it follows that ***l*** is non-self-plotting. QED.

So her token of NON-SELF-PLOTTING LOCATION has a location that is both self-plotting and not. What this means is that plotting the locations of her location-denoting tokens is an impossible task. For in a context, a token of NON-SELF-PLOTTING LOCATION ends up having a contradictory location. Thus, if Keona uses such a token to judge the locations of location-denoting tokens, any apparent success rests on her failing to recognize that *that very token*—one occurring in *each* such judgment—*could not have any possible location.* Or more precisely, the location of any such token cannot be *consistently represented in her model.*

Hence, assuming inconsistency is not an option, a map/theory that aims to represent *every* location of its own symbols cannot use the notion of a "non-self-plotting location" or an equivalent. But then, some location-facts will go unrepresented. More broadly, due to the heterologico-location paradox, we cannot have a consistent theory about the location of all our location-representations. *In particular, we cannot have a consistent theory that locates all location-representations inside the skull.*[28]

## *8. Closing.*

My hope is that the foregoing exemplifies a kind of endeavor which is distinctly philosophical, and which contributes something important. We started with a puzzle, the Place-Smart argument, and eventually explained it as an attempt to cross the limits of what can be consistently represented. Specifically, we identified certain inconsistencies that arise when trying to represent our own representations, as part of the same domain they represent. And I take these paradoxes not just to be minor kinks in our theoretical edifice. They instead represent important and inevitable limits on what can be accomplished. (The map analogy is better at making this clear, though the location paradox is more convincing that the problem is genuine.) Further, if the reader has not already noticed, the issues are not unique to perceptual representations or qualia. Mental representation, of any sort, will give rise to similar problems. But theory dualism is a way to regiment our commitments about mental representation, in a way that avoids the paradoxes (and without any codes).

When Russell discussed the matters like the heterological paradox, he too recognized that they were outgrowths of deeper and more generally problematic situations. This is why his remedy was not merely to forbid a few outlier terms like 'heterological'. Rather, he recognized that a more wide-ranging, systematic solution was called for. The remedy came in the form of the "Vicious Circle Principle." Russell stated the VCP in a few different ways, including:

(a) "Whatever involves *all* of a collection must not be one of the collection" (1908/1956, p. 63).

(b) "If, provided a certain collection had a total, it would have members only definable in terms of that total, then the said collection has no total" (1910/1967, pp. 31, 37).

The applicability of these ideas to the map example should be apparent, and they are just as applicable to the physicalist's model of the world. The regresses owe to one of a totality being defined by the totality itself. (The nested map is defined by the unnested map of which it is a part.) But because of that, the part is never realized *in toto*, and so the whole itself is never realized *in toto* ("the said collection has no total"). From the assumption that the whole itself *is* well-defined, we are thus able to derive contradictions. Correlatively, if we enforce the VCP, we

[28] This should not be read as confirming externalism about representation (Burge 1979; Hurley 2002). If anything, it vindicates "nowhere-ism." Plotting our location-representations anywhere breeds contradiction, for similar reasons.

will not engender the regresses, and we will not create contradiction in assuming that the complete map or a complete physicalist model exists.[29]

## References

Azzouni, J. (2004). *Deflating Existential Consequence: A Case for Nominalism*, New York: Oxford University Press.

____. (2006). *Tracking Reason: Proof, Consequences, and Truth.* Oxford: Oxford University Press.

____. (2010). *Talking about Nothing: Numbers, Hallucinations, and Fictions*. Oxford: Oxford University Press.

Balog, K. (2012). Acquaintance and the Mind-Body Problem. In C. Hill & S. Gozzano (eds.), *New Perspectives on Type Identity: The Mental and the Physical*. Cambridge: Cambridge University Press.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy* 4: 73–121.

Butcharov, P. (1980). Adverbial Theories of Consciousness. In P. French, T. E. Uehling, & H. Wettstein (eds.), *Midwest Studies in Philosophy, Vol. V: Studies in Epistemology*. Minneapolis : University of Minnesota Press.

Chisholm , R. (1957). *Perceiving*. Ithaca: Cornell University Press.

Dennett, D. (1988). Quining Qualia. In A. Marcer & E. Bisiach (eds)., *Consciousness in Contemporary Science*. Oxford: Oxford University Press.

____. (1991). *Consciousness Explained*. Boston: Little, Brown.

____. (2013). On a Phenomenal Confusion about Access and Consciousness. Lecture at the 5th Online Consciousness Conference. Video available at: http://consciousnessonline.com/2013/02/15/on-a-phenomenal-confusion-about-access-and-consciousness/.

Hofweber, T. (2009). Ambitious yet Modest Metaphysics. In D. Chalmers, D. Manley, and R. Wasserman, *Metametaphysics: New Essays on the Foundations of Ontology* (pp. 260–289), Oxford: Clarendon Press.

[29] The VCP was contentious in Russell's time, since it apparently ruled out some classical mathematics. The axiom of reducibility was supposed to fix this, though the axiom seems *ad hoc* and leads to other difficulties. But the VCP "ruled out" some classical mathematics only in the sense that the VCP *plus* the classical logic of *PM*, were incompatible with some such mathematics. Yet it may just be a logicist prejudice to assume that 'classical' in 'classical mathematics' is a telling allusion to classical logic. As far as I can tell, it is an open question whether classical logic is the logic of such mathematics—as opposed to (say) a paraconsistent logic. I realize this is controversial however. But my point is just that the VCP may not be the dead duck it is often thought to be. See also Smith (ms.) for further defense of the VCP.

Hurley, S. (2002). *Consciousness in Action*. Cambridge, MA: Harvard University Press.

Jackson, F. (1977). *Perception.* Cambridge: Cambridge University Press.

____. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127–136.

Fodor, J. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

____. (2009). *LOT2: The Language of Thought Revisited.* Oxford: Clarendon Press.

Gertler, B. (2001). Introspecting Phenomenal States. *Philosophy and Phenomenological Research*, 63: 305-328.

Kim. J. (1992). Multiple Realizability and the Metaphysics of Reduction. *Philosophy and Phenomenological Research*, 52(1): 1-26.

Kriegel, U. (2009). *Subjective Consciousness: A Self-Representational Theory*. Oxford: Oxford University Press.

Ladyman, J. & Ross, D., w/ Spurrett, D. & Collins, J. (2007). *Everything Must Go: Metaphysics Naturalized.* Oxford: Oxford UP.

Lewis, D. (1986). *On the Plurality of Worlds*. Malden, MA: Blackwell.

Lycan, W. (1979). "The Trouble with Possible Worlds," in M. Loux (ed.), *The Possible and the Actual* (pp. 274–316), Ithaca, NY: Cornell University Press.

____. (1987). *Consciousness*. Cambridge, MA: MIT Press.

____. (1996). *Consciousness and Experience*. Cambridge, MA: MIT Press.

____. (2014). The Intentionality of Smell. *Frontiers in Psychology* 5: 1–8.

McGinn, C. (1989). Can We Solve the Mind-Body Problem? *Mind* 98: 349.

Molyneux, B. (2011). On the Infinitely Hard Problem of Consciousness. *Australasian Journal of Philosophy* 89 (2): 211–228.

Parent, T. (2008). Quine and Logical Truth. *Erkenntnis* 68(1): 103–112.

____. (2013). In the Mental Fiction, Mental Fictionalism is Fictitious. *The Monist* 96(4): 608–624.

____. (ms.) Conservative Meinongianism, draft available at http://www.unc.edu/~tparent/conservativeMeinong.pdf

Place, U.T. (1956). Is Consciousness a Brain Process? *British Journal for Psychology* 47: 44–50.

Rosenberg, J. (1998). *Thinking Clearly about Death*, 2nd edition. Indianapolis: Hackett.

Russell, B. (1908/1956). Mathematical Logic as Based on the Theory of Types. *American Journal of Mathematics* 30: 222–262. Pagination is from the reprint in R. Marsh (ed). *Logic and Knowledge* (pp. 59–102), London: Allen and Unwin.

Russell, B. & Whitehead, A.N. (1910/1967). *Principia Mathematica, Vol. 1*. Cambridge: Cambridge University Press. Pagination is from *Principia Mathematica to 56**. Cambridge: Cambridge University Press.

Sellars, W. (1960). Being and Being Known. *Proceedings of the American Catholic Philosophical Association* 35.

____. (1962). Philosophy and the Scientific Image of Man. In R. Colodny (ed.), *Frontiers of Science and Philosophy.* Pittsburg: U. of Pittsburg Press.

____. (1963). Phenomenalism. In his *Science, Perception and Reality*. London: Routledge and Kegan Paul.

Smart. J.J.C. (1959). Sensations and Brain Processes. *Philosophical Review* 68: 141–156.

Smith, P. (ms.). Induction and Predicativity. Unpublished draft, retrieved 10 June 2014 from http://www.logicmatters.net/resources/pdfs/InductionAndPredicativity2.pdf.

Tye, M. (2009). *Consciousness Revisited: Materialism Without Phenomenal Concepts*. Cambridge, MA: MIT Press.

Wittgenstein, L. (1921/1961). *Tractatus Logico-Philosophicus*. D. Pears & B. McGuinness (trans.) London: Routledge and Kegan Paul.