GEORGE BEALER

THE SELF-CONSCIOUSNESS ARGUMENT

Why Tooley's Criticisms Fail

(Received 17 April 2000)

ABSTRACT. Ontological functionalism's defining tenet is that mental properties can be defined wholly in terms of the general pattern of interaction of ontologically prior realizations. Ideological (or nonreductive) functionalism's defining tenet is that mental properties can only be defined nonreductively, in terms of the general pattern of their interaction *with one another*. My Self-consciousness Argument establishes: (1) ontological functionalism is mistaken because its proposed definitions wrongly admit *realizations* (vs. mental properties) into the contents of self-consciousness; (2) ideological (nonreductive) functionalism is the only viable alternative for functionalists. Michael Tooley's critique misses the target: he offers no criticism of (1) – except for an incidental, and incorrect, attack on certain self-intimation principles – and, since he himself proposes a certain form of nonreductive definition, he tacitly accepts (2). Finally, as with all other nonreductive definitions, Tooley's proposal can be shown to undermine functionalism's ultimate goal: its celebrated materialist solution to the Mind-Body Problem. The explanation of these points will require a discussion of: Frege-Russell disagreements regarding intensional contexts; the relationship between self-consciousness and the traditional doctrine of acquaintance; the role of self-intimation principles in functionalist psychology; and the Kripke-Lewis controversy over the nature of theoretical terms.

1. "SELF-CONSCIOUSNESS"

At the outset of my paper "Self-consciousness" I distinguished two theories that go by the name 'functionalism'. First, there is ontological functionalism, whose defining tenet is: "[M]ental properties can be defined wholly in terms of the general pattern of causal (or functional) interaction of ontologically prior 'realizations' and so in this sense are second-order" (p. 69).[1] Second, there is ideological functionalism, which abandons the defining tenet of ontological functionalism but which nevertheless maintains that the standard mental properties "can at least be nonreductively identified in terms of the general pattern of their interaction *with one another*" (p. 73).

Philosophical Studies **105:** 281–307, 2001.

The list of ontological functionalists is long. It includes pretty much all the "Australian functionalists" – Lewis, Armstrong, recent Jackson, Pettit, and, surprisingly, Chalmers (on the propositional attitudes). And it includes most of the "American functionalists" – early Putnam, early Fodor, Block (on the propositional attitudes), Shoemaker (until his shift to ideological functionalism in response to "Self-consciousness"), Loar, Harman, Rey and many others.

My goal in the paper was to establish two main theses. First, "self-consciousness constitutes an insurmountable obstacle to ontological functionalism" (p. 69). The problem may be put as a dilemma. Either the ontological functionalist's definitions "would require the wrong sorts of things to be the contents of self-consciousness: the contents would have to be propositions involving these 'realizations' rather than the mental properties themselves" (p. 69). Or else the right-hand sides of such "definitions" contain undefined psychological expressions, in which case ontological functionalism would fail for that reason (see section 1.2.4 "Alternative Treatments of $\mathcal{P}$" and also note 18). My second thesis was that the only way out of the problem is to *revise* the ontological functionalist's definitions. In these revised definitions the standard mental properties need to be "nonreductively identified in terms of the general pattern of their interaction *with one another*" (p. 73). In other words, the only way out of the problem is to retreat to ideological functionalism. But this means that the resulting nonreductive definitions would "violate the primary tenet of ontological functionalism, namely, that the standard mental properties be definable wholly in terms of the general pattern of causal (or functional) interaction of ontologically prior 'realizations' " (p. 105). These nonreductive definitions would instead endow the standard mental properties "with an ontological primacy inconsistent with the basic functionalist picture" (p. 105). What makes this shift so significant is that it undermines ontological functionalism's most celebrated payoff, namely, a materialist explanation of the relationship between our physical and mental properties and, in turn, a materialist solution to the Mind-Body Problem.

2. MICHAEL TOOLEY'S RESPONSE

In "Functional Concepts, Referentially Opaque Contexts, Causal Relations, and the Definition of Theoretical Terms,"[2] Michael Tooley (hereafter MT) offers a bold and intriguing challenge to the argument of "Self-consciousness." In the present paper I show that this challenge fails and that my theses and arguments are entirely untouched. Besides setting the record straight, there is a special reason for explaining the reasons in detail. A number of other people have also voiced concerns similar to MT's:[3] it is time to get to the bottom of the matter. Further, as a result of MT's commentary several interesting auxiliary issues inevitably arise along the way: the Frege-Russell debate over the best treatment of intensionality (section 3); the relationship between self-consciousness and acquaintance (section 3); the role of self-intimation principles in functionalist psychology (section 5); the Kripke-Lewis controversy over the nature of theoretical terms (section 6).

The primary reason MT's challenge (and the other similar challenges) fails is that it misses the larger dialectic of the argument. Specifically, it takes me to be advocating, not the above stated pair of theses, but rather a single, very different thesis, namely, that *all* forms of Ramsified functional definitions of mental properties are undermined by the phenomenon of self-consciousness. On this construal, if one were able to construct even one form of Ramsified functional definition not undermined by the Self-consciousness Argument (including even the nonreductive ideological definitions which I say do avoid the argument), I would be refuted. But this construal is plainly mistaken, as the above quotations make clear. By virtue of mistaking the target in this way, MT's challenge does nothing to endanger either of my two theses: self-consciousness is indeed a fatal obstacle to ontological functionalism, and this obstacle can be avoided only by renouncing ontological functionalism's primary ontological tenet and therewith its original materialist ambitions. In rough outline, my reply to MT will go as follows.

Thesis (1). My argument against ontological functionalism (pp. 77–80) is really a proof. MT evidently accepts that the proof is valid. Regarding its soundness, he questions two things. First, he questions whether it is appropriate for ontological functionalists

to include various self-intimation principles in the psychological theory upon which their Ramsified definitions are based. His argument for this is unsound, as I will show in section 5. What is more, leading ontological functionalists themselves (e.g., Lewis and Shoemaker; see quotations below) explicitly advocate the inclusion of such principles. In any event, since his main challenge lies elsewhere, MT is willing for the sake of argument to assume that the inclusion of such principles is acceptable. Second, MT suggests that one of the premises in the proof involves an improper treatment of intensional contexts. But, not only is there no improper treatment of intensionality on my part, the suggestion that there is one reveals a crucial, though natural, misunderstanding of the structure of the argument. It is ontological functionalist (not I) who need a way to Ramsify self-intimation principles. But the Ramsified definitions available to them either lead to the absurdity described above or they contain undefined psychological expressions and so fail for that reason. In this way, MT produces no sound objection to my argument against ontological functionalism. Thesis (1) is thus left entirely intact.

MT's discussion of my argument against ontological functionalism might give one the impression that I myself advocate giving Ramsified definitions in the reductive style to which ontological functionalists are committed. But this is certainly wrong. After all, my goal was to show that ontological functionalism is untenable precisely because its reductive functional definitions yield the wrong results. So I obviously would not myself advocate such definitions. (Similarly, since I am not committed to any particular style of Ramsified definition in the case of MT's hypothetical property of being "detectably-water-soluble," his discussion of that property has no bearing on my argument. See section 6 for more on this example.) I do, however, say what Ramsifying functionalists must do to avoid the problem that besets ontological functionalism: they have no choice but to abandon reductive functional definitions and turn to the nonreductive definitions of ideological functionalism. This conclusion is just the first part of my second thesis.

Thesis (2). As with thesis (1), MT's paper leaves my second thesis entirely intact. To begin with, MT evidently agrees with my claim that, if someone wants to give successful Ramsified definitions of

mental properties, *revision* is in order; after all, a good part of his paper is devoted to developing his own revised style of Ramsified definition. More significantly, the style of definition he eventually proposes is in effect just an instance of the same general style of definition I claim is inevitable.[4] In short, MT's positive proposal is actually in agreement with Thesis (2).

I claim that successful Ramsified definitions must employ a type of variable whose intended value is the very mental *property* being defined. MT's proposed definitions employ instead a type of variable whose intended value is the very mental *concept* being defined – where that mental concept is trivially *necessarily equivalent* to the associated mental property. (A concept and property are necessarily equivalent iff, necessarily, the concept applies to an object iff the object has the property. Let it be granted here and at relevant points below that there is a cogent distinction between properties and concepts.) Just as with the essentially simpler nonreductive definitions I focus upon in the paper, MT's nonreductive definitions violate ontological functionalism's defining tenet (that "mental properties can be defined wholly in terms of the general pattern of causal (or functional) interaction of ontologically prior 'realizations' "). This is why MT's definitions, like all other nonreductive definitions, undermine functionalism's elegant explanation of the relationship between our physical and our mental properties and, in turn, its solution to the Mind-Body Problem. MT's proposal thus does nothing to protect functionalism from the destructive effects of the Self-consciousness Argument.

In what follows I will spell out these points in greater detail. As I have indicated, this should be of value, not just to set the record straight, but also because a number of other people have had similar responses to the Self-consciousness Argument. In addition, the discussion will, I hope, be of intrinsic interest, as is MT's own positive proposal about how best to formulate Ramsified definitions.

3. THESIS (1) AND REDUCTIVE FUNCTIONAL DEFINITIONS

Let $\mathcal{A}$ be the psychological theory upon which ontological functionalists wish to base their Ramsified definitions. Let A result from $\mathcal{A}$ by replacing psychological predicates with associated predicate

variables 'R_1', 'R_2', ... Let '**R**' be short for 'R_1, R_2, ...'. Then, assuming that 'is in pain' is the first psychological predicate occurring in $\mathcal{A}$ and 'thinks' the second, ontological functionalists then propose the following standard functional definitions:

x is in pain iff_{def} there exist first-order realizations **R** satisfying A and x has R_1.
x thinks q iff_{def} there exist first order realizations **R** satisfying A and x is related by R_2 to q.

And so on for other standard mental properties and relations, including the relation of being self-consciously aware.

Conscious mental properties and relations, including the relation of being self-consciously aware, are characterized by a number of quite distinctive interactive principles. For example, the following self-intimation principle $\mathcal{P}$: if a person is in pain and engaging in introspection, he will be self-consciously aware that he is in pain. Perhaps qualifiers need to be added – for example, 'is in pronounced pain', 'is engaging in thorough and attentive introspection', '*ceteris paribus*', 'probably' (for alternatives to $\mathcal{P}$, see section 5 and note 25 below). The point is that some such principles, with or without qualifiers, should belong to psychological theory $\mathcal{A}$, given that $\mathcal{A}$ is comprehensive. Leading ontological functionalists agree. For example, David Lewis says, "[Functionalism] allows us to include other experiences among the typical causes and effects by which an experience is defined. It is crucial that we should be able to do so in order that we may do justice, in defining experiences by their causal roles, to the introspective accessibility which is such an important feature of any experience. For the introspective accessibility of an experience is its propensity reliably to cause other (future or simultaneous) experiences directed intentionally upon it, wherein we are aware of it."[5] Echoing much the same point, Sydney Shoemaker tells us, "[I]n many cases it belongs to the very essence of a mental state (its functional nature) that, normally, its existence results, under certain circumstances, in there being such awareness of it."[6]

For simplicity, suppose that $\mathcal{A}$ is a conjunction of some complex clause $\mathcal{B}$ and $\mathcal{P}$. B results from $\mathcal{B}$ by replacing psychological predicates with predicate variables as before. Assume that 'introspects' and 'is self-consciously aware' are, respectively, the third and fourth psychological predicates occurring in $\mathcal{A}$. Then, it would

seem that the ontological functionalist's recipe for forming functional definitions would yield the following definition:

x is self-consciously aware that P iff$_{def}$ there exist first-order properties **R** such that (i) they satisfy B; (ii) if x is R1 and R_3, then x will be related by R_4 to the proposition that he is R1; (iii) x is related by R_4 to the proposition that P.

(Note that clause (ii) results from $\mathcal{P}$ by replacing occurrences of 'is in pain' with 'R_1', 'introspects' with 'R_3', and 'is self-consciously aware' with 'R_4'. As I will explain in a moment, my larger argument does not depend on dealing with $\mathcal{P}$ in this way. I happen to believe, however, that this or something equivalent to it is what the ontological functionalists' standard recipes in the published literature require them to do).[7]

This definition implies, however, that propositions involving first-order realizations of the property of being in pain would be included as typical contents of one's self-conscious awareness. Ontological functionalism tells us that, if x is in pain and engaging in introspection, then x is R_1 and x is R_3, where R1 and R_3 are first-order realizations of the property of being in pain and the property of engaging in introspection, respectively. This and clause (ii) imply that in such a circumstance x will be related by R_4 to the proposition that he is R_1. But the definition of the relation of being self-consciously aware tells us that, if x is related by R_4 to an arbitrary proposition P, then x is self-consciously aware that P. Therefore, in the envisaged circumstance x will be self-consciously aware that he is R_1. But R_1 is not the property of being in pain, but rather a first-order realization, say, the property of having firing C-fibers. The upshot is that the definition admits the wrong sorts of things into the contents of self-conscious awareness.[8]

Many commentators (evidently including MT) suggest that this argument involves an improper treatment of intensional contexts, specifically, in the above Ramsification of principle $\mathcal{P}$. After all, so the suggestion goes, the embedded occurrence of 'is in pain' is an intensional occurrence. Two points are in order. First, this suggestion seriously is mistaken as a point of intensional logic. This is immediately evident in the context of Russellian intensional logic, which is the framework I was explicitly using when I originally presented the argument (see note 14).[9] In a Frege-Church intensional logic, ontological functionalists would need to deal with the

embedded occurrence of 'is in pain' in a somewhat more complicated way, but the resulting definition is subject to virtually the same argument. (See below. A thorough refutation of the intensionality objection is given in my "Ramsification and Intensionality.") Second, and more importantly, the intensionality objection betrays a misunderstanding of the dialectic of the larger argument. Before I explain why, it should be noted that the entire issue of intensionality may be bypassed by a simple change of example.

Consider the psychological relation of self-attribution (which Roderick Chisholm and David Lewis would have us focus upon). On analogy with principle $\mathcal{P}$, $\mathcal{A}$ would contain the following self-intimation principle: if x has the property of being in pain and x has the property of introspecting, then x will self-attribute the property of being in pain (or, more coloquially, x will attribute the property of being in pain to himself). Here the occurrences of the singular term 'the property of being in pain' following 'has' in the antecedent and following 'self-attribute' in the consequent are both plainly extensional. (For example, we could replace *salva veritate* both occurrences of the singular term 'the property of being in pain' with occurrences of any co-designating expression, say, 'the mental property most often used as an example in philosophy of mind'.) So, uncontroversially, the Ramsification of this principle is: if x has R_1 and x has R_3, then x is related by R_5 to R_1.[10] Then the rest of the Self-consciousness Argument goes through *mutatis mutandis*. The absurd conclusion follows, namely, that it is commonplace for ordinary persons x to attribute to themselves first-order realizations of the property of being in pain (say, the property of having firing C-fibers), rather than the property of being in pain itself.[11]

A similar point may be made if 'is self-consciously aware that' is replaced with 'is self-consciously aware of' in the formulation of principle $\mathcal{P}$: if x has the property of being in pain and the property of engaging in introspection, then x will be self-consciously aware of (the state of affairs of his) being in pain. MT grants that states of affairs are analyzable in terms of properties (e.g., being in pain, etc.) and perhaps individuals. He also grants that these analyses may be given entirely in extensional language. Therefore, if he were to Ramsify this formulation of principle $\mathcal{P}$ in conformity with the defining tenet of ontological functionalism (that mental properties

be definable wholly in terms of the pattern of interaction of first-order realizations), he would have no choice but to do so in a way that falls prey to the Self-consciousness Argument.

Now let us consider the intensionality objection directly. The objection is that the argument involves an improper treatment of intensionality in connection with the Ramsification of principle $\mathcal{P}$. As noted, however, this objection betrays a crucial (though natural) misunderstanding of the dialectic of the larger argument.

It is the ontological functionalist (not I) who must tell us how to construct their Ramsified definitions and, in particular, how to Ramsify principle $\mathcal{P}$ (and kindred principles). In my original statement of the argument (summarized above) I simply followed what I take to be the ontological functionalists' standard recipe for Ramsifying (see note 8 above). In making this supposition I was trying to be faithful to their stated intentions in the published literature. But my larger argument takes no stand on this matter. Why? Because it is the ontological functionalist who needs *some* method for Ramsifying $\mathcal{P}$ that yields the correct results and that is consistent with the defining tenet of ontological functionalism. Therefore, given that the method of Ramsifying $\mathcal{P}$ used above does not yield the correct results, ontological functionalists need an alternate method that meets these two conditions. Do they have one? This brings us to the other horn of the dilemma posed by Thesis (1).

One proposal which ontological functionalists might try to make is simply to leave embedded occurrences of psychological expressions untouched. But in this case their resulting "definitions" would have psychological expressions occurring on the right-hand sides and so would not count as acceptable definitions. (See section 1.2.4 "Alternative Treatments of $\mathcal{P}$.") Whether this results in a true circle may be debated. What is certain is that ineliminable psychological expressions would occur on the right-hand sides of (at least some of) these definitions. Hence, the ontological functionalist's definability thesis would be defeated. This is the second horn of the fatal dilemma that the Self-consciousness Argument creates for ontological functionalism.[12]

Do ontological functionalists have any other response besides the above two? In particular, besides a variable (like 'R_1') whose range is restricted to first-order properties, is there any other sort

of variable which ontological functionalists may substitute for the embedded occurrence of 'is in pain'? Trivially, no. For using any other sort of variable would violate the defining tenet of ontological functionalism – that the standard mental properties may be defined wholly in terms of the general pattern of interaction of first-order properties.

Of course there is one candidate sort of variable that, although strictly in violation of this tenet, would nevertheless be in the spirit of ontological functionalism. I have in mind replacing 'is in pain' with a variable whose range is restricted to first-order realizer concepts, rather than first-order realizer properties. But this is no advance at all. For this proposal would lead to an obvious variant of the first prong of the dilemma: the resulting definitions would wrongly imply that the typical contents of our self-conscious awareness would be propositions involving, not the standard mental concepts (e.g., the concept of being in pain), but rather first-order realizer concepts which are necessarily equivalent to the first-order physical properties which realize the standard mental properties (e.g., a first-order concept that is necessarily equivalent to the concept of having firing C-fibers). Once again, an absurd result.

The upshot is that the Self-consciousness Argument still foils ontological functionalism. MT provides no reason to doubt this (except his doubt about relevant self-intimation principles, which I will deal with in section 5). So far, therefore, the Self-consciousness Argument is on track.

The intensionality challenge, however, raises a general question about what logical framework to use when dealing with intensional contexts: a broadly Russellian property-based intensional logic or a broadly Fregean concept-based intensional logic. MT gives an interesting new *metaphysical* argument (vs. some logico-linguistic argument, as is the norm) aimed to show that the former approach is defective and that only the latter can be adequate (pp. 261f.). Upon closer examination, however, this argument is seen to turn on an equivocation in the use of 'property': MT uses it narrowly to apply only to basic properties whereas Russellian intensional logicians use it to apply to nonbasic properties as well. Among the latter are Russellian complex properties, that is, fine-grained properties having a logical form (in the same way Russellian propositions have

logical form).[13] In this connection, Russellians hold that complex verb phrases express associated complex properties. Now given that the goal of MT's argument is to attack Russellian logical theory, he must use 'property' in the way it is used in the context of that theory if he wishes to hit his target. But when 'property' is used this way, a key premise of MT's argument is seen to be false, namely, the premise that properties that cannot have instances cannot even exist. While this might be true for basic properites, it is not true for complex properties, for example, the complex property expressed by the complex verb phrase 'is the largest prime'. The first theorem of number theory concerns this property and teaches us *it* is a property which no number can have. Thus, when 'property' is used unequivocally in this manner, MT's argument does nothing to call into question Russellian intensional logic (nor the associated formulation of the Self-consciousness Argument). To do so, MT would need an independent argument that Russellian complex properties cannot exist.

I hold that, when properly systematized, the Russellian picture contains an array of fine-grained intensional distinctions capable of doing all the legitimate work that can be done by those in the Fregean picture.[14] What, then, survives of the contest between the Russellian and Fregean approaches to intensional logic? Nothing significant except that Fregeans hypothesize an excessively complex syntax for our ordinary intensional constructions. For this reason, the Russellian approach prevails on grounds of simplicity. (But to promote fruitful engagement with MT, I will endeavor throughout the present paper to follow MT in speaking as though there exists a realm of concepts above and beyond the realm of properties.)

There is a certain irony associated with the intensionality objection. Early in his paper (p. 253), MT tells us that "acquaintance plays no part in [Bealer's self-consciousness] argument." But the Self-consciousness Argument is in fact very much concerned with *acquaintable properties* (i.e., self-intimating properties such as being in pain, feeling giddy, etc.). According to the traditional doctrine of acquaintance (to which MT voices no objection), we can be directly aware of such properties without any mediation, conceptual or otherwise. If this is right, then even if MT's mediating concepts were required for the analysis of other sorts of mental

contents, they would be inappropriate to the contents with which the Self-consciousness Argument is concerned.[15] Three things follow. First, insofar as the Self-consciousness Argument is concerned with acquaintable properties, it is impossible for it to be guilty of wrongly bypassing concepts. Second, MT's view itself (that, in bypassing concepts, the argument involves an intensionality error) involves an intensionality error of its own, for it is committed to inserting a layer of mediating concepts where there is none (thereby wrongly identifying the contents of self-conscious awareness). Third, this intensionality error on MT's part creates a corresponding error in his new style of Ramsified definition (at least) in the case of the self-consciousness relation.

4. THESIS (2), NONREDUCTIVE FUNCTIONALISM, AND THE MIND-BODY PROBLEM

It is clear what Ramsifying functionalists must do to avoid the problem created by the phenomenon of self-consciousness. They need to *revise* the ontological functionalist's Ramsified definitions in one of two ways. Either the embedded occurrence of psychological expressions should be replaced with predicate variables whose intended values are not first-order realizations but rather the standard mental properties themselves. Or they should be replaced with variables whose intended values are the standard mental concepts (assuming as I am throughout this paper that there is a cogent distinct between properties and concepts). Either way, it is clear that we would have a deep violation of the defining tenet of ontological functionalism. We would instead have a case of ideological functionalism. Since in MT's positive proposal embedded psychological predicates are replaced with predicate variables whose intended values are the standard mental concepts themselves (the concept of being in pain, etc.), it is an example of the second alternative, and so it amounts to a complicated form of ideological functionalism. For this reason, MT's proposal does not rescue ontological functionalism from the Self-consciousness Argument.

The envisaged revised definitions are *nonreductive* in the sense that, unlike the ontological functionalist's definitions, they do not equate mental properties (concepts) with second-order constructions

wholly from the general pattern of ontologically prior first-order realizations. On the contrary, these definitions endow mental properties (mental concepts) with an ontological primacy inconsistent with the standard functionalist picture: mental properties (concepts) are now taken to be antecedently given ontological primitives already there waiting to constitute the content of our thought. Therefore, these definitions, at most, simply locate mental properties (concepts) within the space of ontologically primitive properties (concepts).

In "Self-consciousness" (section 2.3) I discussed in some detail the simplest sort of nonreductive functional definition. Such definitions take the following form: to be in pain is to have a property that plays the "pain-role" in psychological theory $\mathcal{A}$; to think that P is to be related to P by a relation that plays the "thinking-role" in psychological theory $\mathcal{A}$. And so forth. More formally,

x is in pain iff_{def} there exist properties **R** satisfying A and x has R_1.
x thinks p iff_{def} there exist properties **R** satisfying A and x is related by R_3 to p.

The intention is that in each such definition the standard mental properties themselves are to be among the values of the variables ('R_1', etc.) occurring within the right-hand side.[16] In this way the standard mental properties are being defined in terms of their interaction with themselves and one another, for obviously the standard mental properties are themselves satisfiers of A. It is this feature that opens up the possibility of getting right the contents of our everyday self-conscious thoughts. At the same time, advocates of such definitions commit themselves to the thesis that the matrix A can and shall be so restrictive that it admits no unwanted satisfiers of the sort that wrongly found their way into the contents of self-consciousness.[17] The point of this thesis is to ensure that these definitions do not run afoul of the Self-consciousness Argument.

As I indicated in "Self-consciousness" (note 21; p. 90; and section 2.3), there are more complex styles of nonreductive functional definition besides the simple style just described. Virtually everyone who has proposed a revised functional definition in response to the Self-consciousness Argument has offered some form of nonreductive definition – either in the simple style or in one of these more complex styles.[18] The style of definition proposed by MT is an illustration of one of these more complex styles (specifically, it is a concept-theoretic variant of a suggestion I originally

made in note 21). So MT evidently accepts my thesis that the problem of self-conscious thought forces Ramsifying functionalists to accept some sort of nonreductive definition.

In "Self-consciousness" I suggested that all of these more complex definitions should be eschewed in favor of the simpler ones on the grounds that their added complexity is gratuitous. Specifically, I maintained that it can be shown that the more complex definitions are counterexample-free only if the simpler ones are. This is not to say, however, that all of these more complex definitions are counterexample-free. For example, MT's Ramsified definitions actually stipulate that what has causal efficacy are state of affairs involving, not the standard mental properties themselves, but first-order realizations. But this goes against a central tenet of many leading functionalists (e.g., Shoemaker), namely, that the standard mental properties themselves are directly involved in mental-to-mental and mental-to-physical causation. For these functionalists, the state of affairs (or event) of my being in pain – not the state of affairs of my having some first-order realization – is what causes the state of affairs (event) of my deciding to take an aspirin. This commonsense view of mental causation is ruled out by MT's definitions, which would instead pretty much entail epiphenomenalism. A further difficulty in MT's definitions (e.g., 'the concept of being in pain=$_{def}$ the unique concept C such that, for all x, *if* C applies to x, then for some first-order property I, . . .') is that *every* null concept would vacuously satisfy these conditionals, which make up the entire matrix in these definitions. For this reason, the matrix never has a unique satisfier. Accordingly, the definiens ('the unique concept C . . .') picks out *no* concept at all, and so the definition is necessarily mistaken. A plausible step toward solving this problem would be the following reformulated definition:

> The concept of being in pain = $_{def}$ the unique concept C such that, necessarily, for all x, C applies to x if and only if, for some first-order property I, ...

(These points about uniqueness apply *mutatis mutandis* to MT's definition (p. 266) of "detectable-water-solubility." A third difficulty in MT's style of definition was mentioned at the close of section 2 and a fourth will be discussed in section 6.)

In the remainder of this section what I have to say will not turn on the above claim that various more complex nonreductive functional

definitions should be eschewed in favor of the simple ones (stated earlier) on the grounds that their added complexity is gratuitous. To simplify my presentation, however, it will be convenient to assume for a while that these simple definitions are correct. This simplifying assumption will prove harmless, for it will be evident that my comments would hold even if some more complicated nonreductive definitions were needed (including perhaps those in the spirit of MT's definitions).

Historically, the primary goal of functionalism has been a materialist explanation of the relationship between our physical and mental properties and, in turn, a materialist solution to the Mind-Body Problem.[19] According to this account, a comprehensive description of a being's physical properties, together with the correct functional definitions of relevant mental properties, imply as a purely logical consequence that the being has associated mental properties.[20] For example, consider a being y (e.g., you), and let D be a comprehensive particle-for-particle description of all of y's first-order physical properties at a time when y is in pain. Consider the ontological functionalist definition: y is in pain iff_{def} there exist first-order realizations **R** satisfying A and y has R_1). Then, according to the traditional account, there would be complex predicates in D describing an A-like pattern among first order physical properties, among which is a relevant property had by y; furthermore, there would be a match-up between these complex predicates and corresponding predicate variables in the right-hand side of this definition. As a result of this match-up, the right-hand side of the definition would be a logical consequence of D (by simple existential generalization on the indicated complex predicates in D). Therefore, if the definition is correct, it would follow logically that y is in pain. Assuming that this generalizes, we would have a very elegant materialist account of the relationship between y's physical and mental properties! But we know this account cannot work generally, for the Self-consciousness Argument shows that the ontological functionalist definition of self-consciousness is mistaken. And similar considerations show that the ontological functionalist definition of other mental properties and relations (e.g., conscious thinking) are likewise mistaken.

The natural way to try to save this account is to invoke nonreductive functional definitions, which were adopted precisely to escape the Self-consciousness Argument. But this is no gain at all, for in this new setting it can be shown that D simply cannot have the sort of logical consequences posited in the above functionalist account. Most people are convinced of this as soon as they realize the following. In order for various nonreductive functional definitions to be correct, the matrix A must be so restrictive that it will not be satisfied by the sort of first-order physical properties that were thought by ontological functionalists to be satisfiers of A.[21] Because of this, however, the hoped-for match-up between relevant complex predicates in D and associated predicate variables in A will simply be missing. Hence, the standard logical inference route envisaged by functionalists (i.e., existential generalization on the relevant complex predicates in D) is of no use. Hence, the *grounds* functionalists thought they had for thinking that having this or that mental property is a logical consequence of D plus correct functional definitions are altogether missing. Thus, the above functionalist account of the relationship between our physical and mental properties is wholly unjustified, a dogmatic holdover from ontological functionalism. This constitutes a major epistemic defeat for functionalism's picture of the body-to-mind relationship.

It of course does not follow from the failure of the standard inference route that the functionalist's envisaged logical consequence relationship does not exist. But it turns out that we can show, by means of countermodels, that it indeed does not hold.[22] (This, however, is not the place to present these countermodels; for details, see my "Ramsification and Intensionality.")

The same fate awaits the more complicated nonreductive Ramsified definitions described above (including MT's concept-based definitions). Once again, the hoped-for inference route (existential generalization on relevant complex physical predicates in D) is just missing. Therefore, if functionalists were to adopt any of these Ramsified definitions, the grounds they thought they had for believing in the envisaged logical consequence relationship would again be lost. Moreover, as before, there are countermodels that show that this logical consequence relationship is missing. In this way, these more complicated Ramsified definitions (including

MT's) also undermine functionalism's materialist account of the body-mind relationship and corresponding solution to the Mind-Body Problem.

5. THE VIABILITY OF SELF-INTIMATION PRINCIPLES

Early in his paper (pp. 253–255) MT advances a pair of objections to self-intimation principles like $\mathcal{P}$. If such principles were not viable, the Self-consciousness Argument would collapse. The first of these objections goes as follows:

> The first is that there is reason to believe that there are animals which experience pain, and which have beliefs – including beliefs that they are experiencing pain – but which do not have the capacity for conscious thought episodes. If this is right, then it must be possible to have the concept of being in pain without having the capacity for thought, and therefore it cannot be correct to use [principles like $\mathcal{P}$] in giving an account of what it is to be in pain. (p. 254.)

Three points are in order.

First, as already noted, leading ontological functionalists (Lewis, Shoemaker) are on record as advocating the inclusion of such self-intimation principles in the psychological theory upon which their reductive functional definitions are based. So in my argument against these ontological functionalists, it is they (not I) who assume the appropriateness of this sort of principle. Of course, my wider goal is to refute *all* forms of ontological functionalism. For this purpose I too need to make such an assumption.

Second, suppose for a moment that MT is right to exclude principles like $\mathcal{P}$ from a functional definition of the property of being in pain. Now in his argument, MT takes my target to be the functional definition of being in pain, and he hypothesizes an animal that is in pain, has the concept of being in pain, but lacks the capacity for conscious thought (i.e., the mental state specified in the consequent of MT's substitute self-intimation principle 2). But the target of my actual argument is not the functional definition of the property of being in pain (see p. 78, pp. 88–9) but rather the functional definition of the relation of self-conscious awareness. When MT's example is transferred to the context of my actual argument, the hypothesized animal would therefore need to have self-conscious awareness and the concept of self-conscious awareness and at the

same time lack the capacity for self-conscious awareness (i.e., the mental state specified in the consequent of *my* self-intimation principle $\mathcal{P}$). But this is a contradiction: there cannot be a being who has self-conscious awareness but who lacks the capacity for it![23] Thus, when aimed at the real target, MT's argument depends on the possibility of an example that is impossible. (Incidentally, MT states that nothing turns upon the fact that the relation of self-conscious awareness and kindred relations are the target of my argument and that "we can simplify the argument by assuming that a functionalist account is being offered only for being in pain" (p. 255). Here is a way in which something does turn on this fact; for another see the close of section 3 above.)

Third, even in the case of the functional definition of being in pain (vs. self-consciousness), MT's objection does not go through. For, to make the argument deductively valid, a missing premise needs to be supplied. But it is not clear what this implicit premise could be. I can think of two candidates.

(1) The suppressed premise might be: if it is possible for a creature to have a certain psychological concept (e.g., the concept of being in pain) but to lack a certain cognitive capacity (e.g., the capacity for conscious thought), then no account of the concept may use principles concerned with that cognitive capacity. But not only is this principle implausible, it would, if true, pretty much undermine all functional definitions. To see this, consider a creature that has the concept of believing but that altogether lacks the capacity to make inferences involving, say, disjunctions (although it does have the capacity to make inferences involving, say, conjunctions). Then, when we try to formulate a functional definition of the concept of believing, the envisaged suppressed premise would bar the use of principles concerning inferences involving disjunctions. But there could be a another sort of creature for whom things are just the other way round: the creature has the concept of believing but altogether lacks the capacity to make inferences involving conjunctions (although it does have the capacity to make inferences involving disjunctions). Accordingly, when we try to formulate a functional definition of the concept of believing, the suppressed premise would, in addition, bar the use of principles concerning inferences involving conjunctions. But this is only the

tip of the iceberg. Because there is an endless supply of examples with comparable effect, the suppressed premise would bar just about every relevant psychological principle from the functional definition of the concept of believing, thus dooming such a definition.[24]

(2) The suppressed premise might be: if it is possible for a creature to have concept c but lack the capacity to have concept c', then c' cannot be included in any correct account of c. Again, this premise is false. Here are two counterexamples. It is possible for someone to have the concept of acceleration without having the capacity for the mathematical concept of a second derivative, so the suppressed premise would wrongly imply that the latter concept is barred from a correct account of the concept of acceleration. Similarly, it is possible to have the everyday concept of a calculation without having the capacity for the mathematical concept of a Turing machine; therefore, the latter concept cannot be used in a correct account of the concept of a calculation. Generalizing on these examples, we see that the envisaged premise entails that virtually all interesting conceptual analysis would be impossible.

We come now to MT's second objection to $\mathcal{P}$-like self-intimation principles. Let principle $\mathcal{P}^*$ be: if x is in pain and engaging in introspection, then x will also have the thought that he is in pain. And let $\mathcal{P}^{**}$ be: if x thinks p and is engaging in introspection, then x thinks that he thinks p. MT tells us, "If [$\mathcal{P}^*$] is true, then related principles – including, in particular, [$\mathcal{P}^{**}$] – must also be true" (p. 254). Then he observes that $\mathcal{P}^{**}$ cannot be true, for it generates an unacceptable infinite regress. Hence, by modus tollens, $\mathcal{P}^*$ cannot be true. So goes the objection. The rebuttal goes as follows. First, $\mathcal{P}^*$ simply does not entail $\mathcal{P}^{**}$, so the problem is avoided at the first step. But even if that reply is set aside, there are several alternate principles which resemble $\mathcal{P}^{**}$ but which generate no such regress. Here is an illustration: if x is thinking something and engaging in introspection, then x will think that he is thinking something. (As with principle $\mathcal{P}$, you may add qualifiers if you wish.) In fact, there are many other equally plausible principles in this general family.[25] And the Self-consciousness Argument requires only *one* such self-intimation principle.

6. THE KRIPKE-LEWIS DEBATE ON THE NATURE OF THEORETICAL TERMS

In this final section I will make a few remarks on the technique of Ramsified definitions. This technique (first advocated by R. M. Martin and subsequently advocated by Putnam, Lewis, and Grice independently of one another) is thought by many to be applicable, not just as a method for defining everyday mental properties, but also as a general technique for defining the theoretical terms used in the special sciences (physics, chemistry, etc.). But, as is well known, there is an entirely different approach to theoretical terms, namely, the Kripkean reference-fixing theory of names. On this theory, theoretical terms (as well as most ordinary nontheoretical terms) are often introduced by means of contingent reference-fixing descriptions which provide an unsatisfactory basis for a definition of the term (either an ordinary first-order definition or a Ramsified definition); indeed, those reference-fixing descriptions might be used referentially (vs. attributively) and so might well have a descriptive content that does not even apply to the nominatum (as in Keith Donnellan's martini-man example). And even when the reference-fixing description happens to single out the nominatum attributively, if it were subsequently used as the basis of a definition (either an ordinary first-order definition or Ramsified definition), it would have mistaken modal consequences (e.g., in the meter-stick example, such a definition would wrongly imply that it is necessary that the length of the so and so stick be one meter). And this point generalizes to theoretical terms. On this view, moreover, many such terms can be given a correct definition (specifically, a scientific definition) only after scientific investigation, and most often that definition will be an ordinary first-order definition (e.g., water $=_{\text{def}}$ H_2O) rather than a Ramsified definition based upon the term's behaviour in some scientific theory in which it is embedded. In such cases, the Ramsified definitions would be outright mistaken, for they would have a number of implausible implications.

Here is an example. Suppose that unbeknownst to advocates of a given scientific theory, the theory is satisfied, not just by the intended sequence of theoretical properties, but also by some unintended, even irrelevant, sequence. Then, on the Martin-Putnam version of Ramsification, the resulting Ramsified definitions would wrongly

admit the extensions of those unwanted satisfiers into the extensions of the corresponding theoretical terms. On Lewis's version (which MT evidently endorses), this unwanted outcome does not arise because a unique satisfiability condition is built right into the definition.[26] But in this case an equally undesirable outcome arises: for theories that do not satisfy the uniqueness condition, the associated Ramsified definitions would wrongly imply that the theory's theoretical terms are vacuous. A response (which Lewis hints at, op. cit.) is to hold that any good scientific theory must meet this unique satisfiability requirement. But this requires far too much: we would not deem a scientific theory unsatisfactory simply because it happens to have more than one sequence of satisfiers; moreover, I can see no reason to think that, for every world, uniqueness can always be achieved by the true scientific theories for that world.[27] Finally, with a uniqueness condition built in, these definitions would imply that various laws of nature are necessarily true (even analytic) whereas on the traditional view laws of nature are contingent. None of these implausible consequences besets the Kripke-Putnam picture.

Now MT seems to take as a starting point Lewis's doctrine that the Ramsification method ought to be able to provide correct definitions for any theoretical term. We see this in his effort to give a Ramsified definition of 'detectably-water-soluble' (Instead of this tendentious term, I will use an appropriately neutral primitive expression, say, 'Dwas'.) It seems to me, however, that 'Dwas' fits the Kripkean picture to a tee. In the hypothetical example, it is not introduced by means of a theory but rather by a contingent reference-fixing description. Furthermore, given MT's description of the case, 'Dwas' would have a direct first-order scientific definition (x is Dwas iff_{def} x has such and such molecular structure), just as the Kripkean theory predicts. Assuming (as MT seems to) that this sort of scientific definition would be correct, a Ramsified definition would then seem to be an unnecessary complication. Even more serious, a Ramsified definition would have unwanted modal implications of the sort described a moment ago.[28] For example, it would wrongly imply that the following conditional is necessary (indeed, analytic): if something is Dwas and is immersed in water, it will dissolve. Moreover, if this conditional were necessary, then (given

the correctness of the indicated first-order scientific definition of Dwas) the following conditional would also have to be necessary: if something has such and such molecular structure and is immersed in water, it will dissolve. But, intuitively, neither of these conditionals would be necessary; on the contrary, they would just be contingent laws of nature. At least, this is what the prevailing view on the modal status of laws of nature would say. (MT embraces the contingency of laws of nature elsewhere in his writings.) In this respect, mental properties are very different from the property of being Dwas (assuming that nonreductive functionalism is right), for it is in the very *nature* of mental properties that they interact in the fashion described by, say, self-intimation principles like $\mathcal{P}$. And, intuitively, $\mathcal{P}$ does seem necessary, at least when suitable qualifiers are supplied. (Such principles report our "Cartesian Intuitions," which Sydney Shoemaker, ibid., believes functionalism should preserve.) Here, then, is a central way in which the "detectably-water-soluble" analogy breaks down.

CONCLUSION

The foregoing shows that MT's critique of "Self-consciousness" is unsuccessful. The picture I sought to defend in the paper remains fully intact. (1) Ontological functionalism is undermined by the phenomenon of self-consciousness. (2) The only way to save Ramsified functionalism is to resort to some form of nonreductive functional definition (of which MT's positive proposal is just a complicated instance). Such nonreductive definitions (whether simple or complex), however, undermine the primary goal of functionalism – its materialist account of the body-mind relationship and its solution to the Mind-Body Problem.

NOTES

[1] "Self-consciousness," *The Philosophical Review*, vol. 106, 1997, pp. 69–117. In the paper I criticize two versions of ontological functionalism – the familiar Ramsified version and the Language-of-Thought version. In the present context only my arguments against the Ramsified version are under attack, so I will confine my remarks exclusively to those arguments and the criticisms of them.

Hereafter, unless otherwise indicated, page numbers and note numbers will be those in "Self-consciousness."

[2] *Philosophical Studies*, vol. 105, 2001, pp. 251–279.

[3] For example, during comments and discussion at nearly every presentation of the paper. See also Mark McCullagh, "Functionalism and Self-Consciousness," *Mind and Language*, vol. 15, pp. 500–510.

[4] It is in fact just a concept-theoretic variant of the sort of definition contemplated at the end of note 21 in "Self-consciousness."

[5] P. 103, "An Argument for the Identity Theory," in *Philosophical Papers, Volume I*, New York: Oxford University Press, 1983, pp. 99–107.

[6] P. 59, "The Mind-body Problem," in *The Mind-Body Problem: A Guide to the Current Debate*, Richard Warner and Tadeusz Szubka (eds.), Oxford: Basil Blackwell, 1994, pp. 55–60.

[7] As David Lewis emphasizes, "We must assume that all occurrences of T-terms in the postulate of T are purely referential, open to existential generalization and to substitution by Leibniz's law. We need not assume, however, that the language of T is an extensional language." (p. 80, "How to Define Theoretical Terms," in *Philosophical Papers, Volume I*, New York: Oxford University Press, 1983, pp. 78–95). Now since the embedded psychological predicates in principle $\mathcal{P}$ express acquaintable properties (see section 5), they do occur referentially according to the traditional doctrine of acquaintance. Hence, the Lewis style recipe directs functionalists to replace embedded occurrences of such psychological predicates with predicate variables. (Note that the first step in Lewis's recipe needs to be omitted in the case of embedded predicates. Why? Because if, following Lewis, one were at this first step to replace them with copulas and primitive property names, an intensionality error would arise: the resulting principle would not be a truth of psychology.)

For similar reasons, Putnam's recipe would be effectively the same (section VI, "On properties," in Nicholas Rescher et al., eds., *Essays in honor of Carl G. Hempel*, Dordrecht: D. Reidel, 1970). Likewise for other ontological functionalists in the Putnam tradition.

[8] Another possibility is that there is *no* sequence of first-order realizations **R** satisfying A. If so, the right-hand side of the definition would be null and therefore would certainly not define the relation of being self-consciously aware.

One familiar response to this argument is to "bite the bullet," that is, to hold that propositions involving such first-order realizations really are typical objects of the self-consciousness relation. See my "The Mind-Body Problem" (forthcoming), for an explanation of why this response is mistaken.

Note that in the argument I am talking about the property of being in pain, which is denoted by the canonical gerundive phrase 'being in pain.' All ontological functionalists – "Australian" (e.g., Lewis, et al.) as well as "American" – are committed to identifying this property, not with a first-order realization, but with a second-order functional property (see, e.g., Lewis, ibid.; see also my "Ramsification and Intensionality".)

[9] For Russell, embedded and unembedded predicates do not differ in semantic

value; semantically, they correspond to intensions (formally, propositional functions; informally, properties). In *Principia Mathematica* '(Pain(x) & Introspect(x)) → Self-conscious(x, Pain(x))' is provably equivalent to '(∃f)(f = Pain & [(fx & Introspect(x)) → Self-conscious(x, fx)])'. Here the predicate 'Pain' occurs just once, and no longer within the scope of a psychological predicate. Replacing this occurrence of 'Pain' with predicate variable 'R_1', 'Introspect' with 'R_3', and 'Self-conscious' with R_4, we get '(∃f)(f = R_1 & [(fx & R_3x) → R_4(x, fx)])', which in turn is provably equivalent to '(R_1x & R_3x) → R_4(x, R_1x)'. But the latter is exactly the Ramsification of $\mathcal{P}$ used in the Self-consciousness Argument.

[10] For simplicity I am assuming here that 'self-attribute' is the fifth mental predicate in the psychological theory $\mathcal{A}$.

[11] Against this argument it might be objected that self-attribution is a nonbasic mental relation which is to be defined, not by means of Ramsification, but rather *directly* in terms of the thinking relation itself. (This approach is not available to functionalists, such as David Lewis, who take self-attribution to be definitionally prior to thinking.) On this approach, self-attribution might be defined as follows: for all F, x self-attributes property F iff$_{\text{def}}$ x thinks that he is F. But, to be correct, this definition requires that the embedded occurrence of 'F' on the right-hand side is an externally quantifiable predicate variable ranging over properties. So in this setting the intensionality objection is dead in its tracks.

Some people might hold that we self-attribute concepts, but this is implausible on its face: you self-attribute the *attribute* of being in pain.

[12] Remember, in this present paper we are confining ourselves to Ramsified formulations of ontological functionalism; for simplicity we are supposing that the language-of-thought version is off limits, though that position is dealt with in "Self-consciousness."

[13] Some Russellians hold that, for every complex property, there is a necessarily equivalent simple property. Others disagree, holding that all simple properties are basic properties.

[14] See David Kaplan, "How to Russell a Frege-Church," *The Journal of Philosophy*, vol. 72, 1975, pp. 716–29. Also my *Quality and Concept*, 1982, Oxford and my "A Solution to Frege's Puzzle," *Philosophical Perspectives*, vol. 7, 1993, pp. 17–61.

Someone might try to challenge the claim in the text on the ground that properties are mind-independent entities whereas concepts are mind-dependent. But, not only would this challenge be out of step with Frege's anti-psychologism, a certain modal argument shows that this view of concepts cannot be right. See my "Universals," *The Journal of Philosophy*, vol. 90, 1993, pp. 5–32.

[15] This point draws attention to another slip in MT's argument about intensional logic, namely, its invalid inference from the conclusion that *some* propositions need to be analyzed in terms of concepts (vs. properties), to the further conclusion that *all* propositions need to be analyzed that way.

[16] Thus, unlike the original functional definitions, which were formulated in the predicative logical framework of "ramified type theory" (see Putnam, ibid.), these definitions are formulated in an impredicative type-free logical setting.

[17] Toward this end, the underlying psychological theory $\mathcal{A}$ might contain iterated-attitude clauses such as 'It is possible for someone to think that he is thinking something', and the predicate variables in A might be restricted to natural (or basic) properties. Of course, the thesis in the text would hold if A *implicitly defines* the standard mental properties and so is unqiuely satisfied by them. For in this case, no first-order physical realizations of these properties would also satisfy A.

[18] An exception to this general tendency is the proposal made by Mark McCullagh, ibid.

[19] This goal has been articulated by many people: Putnam, Lewis, Shoemaker, Block, Jackson, and others.

[20] Some Ramsifying functionalists believe that the psychological theory upon which their functionalist definitions are to be based will be *a priori*. For these people, the resulting definitions will also be *a priori*. But other Ramsifying functionalists believe that the relevant psychological theory will be an *a posteriori* scientific theory, and so they believe that the associated definitions will be *a posteriori* scientific definitions. Both sorts of functionalist, however, accept this account.

[21] To illustrate, suppose that the underlying psychological theory $\mathcal{A}$ contains iterated-attitude clauses such as 'It is possible for someone to think that he is thinking something' (see note 17 above). In this case, the associated matrix A would have first-order physical satisfiers only if there were a first-order physical relation r_2 such that it is possible for someone to be related by r_2 to the proposition that he is r_2. But in a molecule-for-molecule description D of our being y, there obviously would be no clause with this logical form; chemical descriptions simply are not like that.

[22] Some functionalists might try to avoid this outcome by trying to stretch how we understand what counts as a physical fact, say, by extending upward the boundary between biology and psychology. For example, suppose that, in the situation contemplated in note 21, y is related by a first-order physical relation r_4 to the proposition that he is in pain. The idea is that some functionalists might propose including this fact among the physical facts. Their hope would be to make the fact that y is self-consciously aware that he is in pain a logical consequence of this and other "phsyical" facts (plus relevant definitions). But functionalists are not free to make this move.

To see why, consider two theses. Thesis I: the Logical Supervenience of the Physical on the Microphysical (i.e., the thesis that, at least in the actual world, the physical facts are logical consequences of the totality of microphysical facts plus any relevant definitions). Thesis II: the Logical Supervenience of the Mental on the Physical (i.e., the thesis that, at least in the actual world, the mental facts are logical consequences of the totality of physical facts plus any relevant definitions). Now every functionalist accepts Thesis I – or at least accepts that it is far more justified than Thesis II. In other words, if they were forced to reject either Thesis I or Thesis II, they would have no choice but to reject Thesis II.

Now, given our standard understanding of what counts as a microphysical fact, the indicated countermodels refute the following Thesis III: the Logical Supervenience of the Mental on the Microphysical (i.e., the thesis that, at least in the actual world, the mental facts are logical consequences of the totality of the microphysical facts plus any relevant definitions). And it would be absurd to try to save this thesis by trying to stretch how we understand what counts as a *microphysical fact*. (Specifically, it would be preposterous to hold that actual facts about the particles and forces constituting y's brain would include a microphysical fact that y is related by the aforementioned first-order physical relation r_4 to the proposition that he is in pain!) In short, our functionalists have no choice but to reject Thesis III. But, taken together, Thesis I and Thesis II, entail Thesis III. Therefore, functionalists have no choice but to reject either Thesis I or Thesis II. But, as we saw above, faced with this choice, functionalists must reject Thesis II. Thus, the ploy of trying to shift upward the biological-psychological boundary is seen to be futile.

[23] Furthermore, the antecedent of principle $\mathcal{P}$ is that x is not only in pain but also engaging in introspection. This of course means *conscious introspection* (is there any other kind?). Surely beings satisfying these two conditions are self-consciously aware that they are in pain (at least if relevant qualifiers are adjoined). MT's example leaves out the requirement that x be engaging in conscious introspection.

[24] Of course, this problem generalizes: functional definitions of just about any psychological concept could be blocked by similar families of examples.

Incidentally, I have heard the following objection to the inclusion of principles like $\mathcal{P}$ (this is not MT's objection). Given that creatures of the sort described by MT are possible, the associated principles would not describe an *essential* property. Hence, such principles should not be included in the theory upon which functional definitions are based. The error, of course, is that such principles are conditionals: *if* . . ., then . . .. A great many (most?) essential properties are associated with such (metaphysically necessary) conditionals.

[25] Here are some examples (i) For propositions p concerned only with phenomenal qualities, if x is thinking p and engaging in introspection, then x will think that he is thinking p. (ii) If x is thinking p and engaging in introspection, then in normal psychological conditions x would be more likely to think that he is thinking p than to think that he is not thinking p. (iii) If x thinks p and considers the question whether he thinks p, then he will think that he thinks p. (The latter principle leads to the usual result but requires a slightly more complicated version of the Self-consciousness Argument. See my forthcoming book *The Integrity of Mind.*)

[26] See pp. 254 ff., "Psychophysical and Theoretical Identifications," *Papers in Metaphysics and Epistemology*, Cambridge: Cambridge University Press, 1999, pp. 248–261. Lewis's proposal rests on Dana Scott's treatment of vacuous descriptions, which I find extremely artificial and unintuitive.

[27] Of course, it is understood here that, e.g., "Electrons are the things to which

our term 'electron' applies" is not a genuine theory in *physics*! And this point generalizes to theoretical terms in other natural sciences.

[28] I am assuming here that the difficulty in MT's definition mentioned in section 4 has somehow been solved.

At several points in his disucssion MT seems to suggest that all the clauses in the theory upon which a Ramsified definition is based would turn out to be 'analytic' if the definition were correct. But there are counterexamples to this claim (e.g. purely existential clauses). It is the case, however, that the claim holds for certain forms of conditionals (of which $\mathcal{P}$ is an illustration).

University of Colorado
Boulder, CO 80302–0232
USA