The Philosophical Review, Vol. 106, No. 1 (January 1997)

Self-Consciousness

George Bealer

Even though most functionalists reject behaviorism and the identity thesis, they view functionalism as the natural successor to these reductionistic views. Their doctrine is ontological: they claim that mental 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.[1] My first purpose is to show that self-consciousness constitutes an insurmountable obstacle to ontological functionalism. Why? Because the envisaged functional 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. The only way to save functional definitions from this obstacle is to expunge the requirement that mental properties be second-order and to accept that they are first-order.[2] But we shall also see that even the resulting "ideological" functionalism, which aims only at conceptual clarification,[3] faces

---

An early version of this paper was given at the CUNY Graduate School in Spring 1985. Recent versions were presented at Arizona State University, Brandeis University, Syracuse University, University of Maryland, University of Miami, and University of Washington. For valuable comments I wish to thank those audiences and Erik Anderson, Jonathan Bennett, Ned Block, David Chalmers, Hartry Field, Mark Hinchliff, Harold Hodes, Stephen Leeds, Brian Loar, Iain Martel, Mark Moffett, Michael Peirce, C. D. C. Reeve, Georges Rey, Stephen Schiffer, Christopher Shields, Sydney Shoemaker, and Ralph Wedgwood.

[1]See, for example, Hilary Putnam, "On Properties," in *Mathematics, Matter and Method: Philosophical Papers,* vol. 1 (New York: Cambridge University Press, 1975), 305–22; Sydney Shoemaker, "Some Varieties of Functionalism," *Philosophical Topics* 12 (1981): 93–119; Brian Loar, *Mind and Meaning* (Cambridge: Cambridge University Press, 1981); Ned Block, "Can the Mind Change the World," in *Meaning and Method: Essays in Honor of Hilary Putnam,* ed. G. Boolos (Cambridge: Cambridge University Press, 1990). In a moment I will say more about the notions of second-order properties and ontological functionalism. Other versions of functionalism will be dealt with in due course; see, for example, note 2.

[2]David Armstrong and David Lewis hold a "functionalist identity theory" according to which mental properties are contingently identical to the first-order "realizations." As we shall see, the argument from self-consciousness can be extended to show that this cannot be right.

[3]In "Method in Philosophical Psychology" (*Proceedings and Addresses of*

an equally insurmountable obstacle. Here the argument—which turns on a feature of self-consciousness evident in Descartes's *cogito*, namely, that it can be "ungrounded"—shows that ideological functionalism fails unless it incorporates the thesis that mental properties are fully "natural" universals. (The argument applies equally against ontological functionalism.) This enables us to reach our final conclusion, namely, that mental properties are *sui generis*: they are first-order, nonphysical, fully "natural" universals. Thus, reductionism in the philosophy of mind, in all of its most promising forms, fails.

Self-consciousness has often been taken as the very mark of the mental, yet functionalists have paid little attention to it. Their focus has been on belief-desire psychology, rational decision theory, and the psychology of sensation, so that they have been largely concerned with "unembedded" psychological phenomena. Likewise, the currently popular objections to functionalism (inverted-spectrum, absent qualia, "what-it's-like," Chinese nation, homunculus head, Chinese room, externalism, anti-individualism, anomalism, utopianism) also focus on "unembedded" psychological phenomena, and each of them, I believe, fails to refute functionalism.[4] In my view, only a change of focus, only a reawakened sense of the centrality of self-consciousness, will enable us to evaluate functionalism as a philosophy of *mind*.

Two points about terminology are in order. First, for convenience I use 'property' for both properties and relations; where relations specifically are relevant, I use that terminology. Second, on the matter of 'first-order' and 'second-order': Ontological functionalists use these terms in the sense of a broadly Russellian ramified theory of properties. According to a simple unramified theory, properties divide into the following hierarchy: properties of individuals, properties of properties of individuals, etc. In a ramified theory each of these divides into an order hierarchy, for example, first-order properties of individuals, second-order properties of individuals, etc. First-order properties of individuals are either primitive properties of individuals or properties of individuals

---

the American Philosophical Association 50 (1975): 23–53), Paul Grice entertains this position (as well as the more familiar ontological functionalism).

[4]I give my reasons in "Mind and Anti-Mind," *Midwest Studies in Philosophy* 9 (1984): 283–328; "Mental Properties," *Journal of Philosophy* 91 (1994): 185–208; and section 1 below.

definable in terms of primitive properties of individuals plus quantification over individuals. Second-order properties of individuals are properties of individuals that are not first-order properties of individuals but are definable wholly in terms of first-order properties of individuals plus quantification over individuals and first-order properties of individuals.[5] It is in this sense that ontological functionalists claim that mental properties are second-order: they are properties of individuals that are not first-order properties of individuals but are definable wholly in terms of first-order properties of individuals plus quantification over individuals and first-order properties of individuals.

This use of 'first-order' and 'second-order' is neutral on the question of the order of mental properties. Russell himself thought that mental properties such as the property of being in pain were primitive first-order properties. By contrast, ontological functionalists hold that whereas various physical properties are first-order, mental properties (including properties like being in pain) are second-order.

The paper is divided into three sections—(1) Ramsified Functionalism, (2) Language-of-Thought Functionalism, and (3) Ideological Functionalism. In section 1, I give a proof that functional definitions of self-consciousness by means of the standard "Ramsification" technique entail that the contents of self-consciousness would wrongly have to be propositions involving first-order "realizations" rather than mental properties themselves. The reason is this: When we describe the general pattern of interaction of the standard mental properties, we find that they function both as

[5]Analogous distinctions hold for relations. See also note 24. Note that in this characterization I used the ontological notion of a primitive property rather than the notion of an undefinable property. The reason is that there is a well-established sense (associated with the project of conceptual clarification) in which some ontological primitives may be said to be definable. For example, someone could say that even though phenomenal colors are ontological primitives, any one of them (for example, Hume's missing shade of blue) is definable in terms of the others (see the close of section 1). Because this is a well-established usage, the term 'ontological functionalism' is preferable to 'definitional functionalism'. Incidentally, in contemporary second-order logic (that part of the simple theory of types known as "the functional calculus of second order"), a property expressed by a formula $\ulcorner A(f) \urcorner$ is often called a second-order property even if the formula A involves no quantification over properties. This usage departs from the original Russellian usage and is not relevant to functionalism.

unembedded properties and as embedded properties. In the latter role, they function as "constituents" of the propositions that are objects of various mental relations. For example, a subject can both be *in pain* and be self-consciously aware that he is *in pain*. When we Ramsify our psychological theory, we are led to quantify both positions—unembedded and embedded. But, according to ontological functionalists, the values of our quantified predicate variables must be first-order "realizations" of mental properties, rather than mental properties themselves. So, in the above example, the value of the embedded predicate variable would be not the property of being in pain, but rather some "realization" of this property (for example, the property of having firing C-fibers). This in turn implies that propositions involving such "realizations" would be the contents of self-consciousness. (The problem generalizes to other multiply embeddable mental relations—thinking, etc.) Within the setting of Ramsified functionalism there are various ways of trying to avoid this unacceptable consequence, but we will see that at best they only defer the problem. Sooner or later it reappears.

What ontological functionalists need to avoid the problem is a systematic way of blocking the offending quantifications of embedded mental predicates, treating the latter as standing only for mental *representations* rather than mental properties. Language-of-thought functionalism is the leading example of this version of functionalism. In section 2, we find that language-of-thought functionalism only puts off the problem. The difficulty resurfaces in connection with Mentalese *psychological* predicates, specifically those whose contents are supposed to be multiply embeddable mental relations (thinking, desiring, self-conscious awareness, etc.). I show that the methods (causal and conceptual role) standardly used by ontological functionalists to fix content are viciously circular in the case of these Mentalese psychological predicates. The only way out of the circle is to retreat to Ramsified definitions, but these all fall prey to some version of the problem of section 1. Thus, both leading versions of ontological functionalism—Ramsified and language-of-thought—fail for the same underlying reason.

In section 3, I take up ideological functionalism, which abandons the ontological claims of ontological functionalism and aims only at conceptual clarification. Given that the standard mental properties cannot be defined in terms of the general pattern of interaction of ontologically prior "realizations," perhaps they can at

least be nonreductively identified in terms of the general pattern of their interaction *with one another*. I show, however, that we can construct a system of nonstandard properties whose general pattern of interaction with one another other matches exactly that of the standard mental properties. This falsifies ideological functionalism as stated (and, as we shall see, ontological functionalism as well). Ideological functionalists might try to save their definitions by building in the requirement that the standard mental properties are "natural" (versus "Cambridge-like"), thereby ruling out the indicated system of nonstandard properties. But this would be to accept our final point—that the standard mental properties are "natural" universals.

1. Ramsified Functionalism

I take the identity thesis to be the thesis that the standard mental properties (for example, the property of being in pain) are identical to first-order physical properties (for example, the property of having firing C-fibers). Most philosophers today reject this thesis—with good reason—and in this paper I will assume that they are right.[6] Ontological functionalists are opponents of the identity thesis; for, unlike identity theorists, they do not believe that mental properties are first-order. Rather, they believe that mental properties are defined by the general pattern of causal (or functional) interaction of ontologically prior "realizations" and are thus second-order. Despite their opposition to the identity thesis, many ontological functionalists nevertheless hold that mental properties are *physical*; this is the most direct legacy of reductionism in philosophy of mind.

The most clear and precise formulations of ontological functionalism are those based on the idea of "Ramsification"—the idea of converting whole theories (sets of sentences) into definitions by replacing "theoretical" predicates with existentially quantified predicate variables.[7] Let $\mathcal{A}$ be a comprehensive description of the

[6]In "Mental Properties" I give my reasons for rejecting the identity thesis.

[7]In "Theories" (in *The Foundations of Mathematics*, ed. R. B. Braithwaite (London: Routledge and Kegan Paul, 1931), 212–36), F. P. Ramsey gave a technique, not for defining "theoretical" terms, but rather for eliminating them by means of existentially quantified predicate variables. To my knowl-

characteristic interaction of the standard mental properties with each other and with physical properties. Suppose that 'is in pain', 'thinks', and so forth are (in order of their first occurrences) the mental predicates occurring in the comprehensive psychological theory $\mathcal{A}$. Let A be the complex expression that results from $\mathcal{A}$ by replacing 'is in pain' with the one-place predicate variable 'R_1', 'thinks' with the two-place predicate variable 'R_2', and so on. Let '**R**' be short for 'R_1, R_2, . . . '. Then, according to this kind of functionalism, the following would be good definitions (remember, I am using 'property' for both properties and relations):

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

And so forth. The idea is that there are first-order properties **R** that function with respect to one another in the way the standard mental properties function with respect to each other; accordingly, these first-order properties may be considered "realizations" of the standard mental properties. So, as ontological functionalism requires, the standard mental properties would be defined wholly in terms of the general pattern of interaction of first-order "realizations" and thus would be second-order.

There is great flexibility in the sort of theory $\mathcal{A}$ may be. Some functionalists inclined to mechanism (for example, early Putnam) would require that $\mathcal{A}$ be (equivalent to) a Turing machine table; others find this requirement too strong. Some (for example, Sydney Shoemaker) would instead reqire that $\mathcal{A}$ be an idealized formulation of commonsense psychology. Others (for example, early Fodor) would instead require that $\mathcal{A}$ be an idealized formulation of scientific psychology. Some (for example, early Putnam) would want the nonpsychological vocabulary in $\mathcal{A}$ to be that of an ideal physics; others (for example, Shoemaker, Grice) would permit the vocabulary of $\mathcal{A}$ to include a wider range of nonpsychol-

---

edge, the idea of using existentially quantified predicate variables to construct the kind of definition described in the text is first found in R. M. Martin's "On Theoretical Constants and Ramsey Constants," *Philosophy of Science* 31 (1966): 1–13. The idea was subsequently taken up by Lewis, Grice, Harman, Loar, Shoemaker, Block, Cummins, Jackson, and many others.

ogical vocabulary (for example, expressions from natural sciences other than physics, "folk" physical and/or teleological properties belonging to no particular science, and perhaps even auxiliary properties from higher mathematics and metaphysics). Some (for example, Loar) might wish to index $\mathcal{A}$ to individuals and times. Some believe that $\mathcal{A}$ would include a normativity predicate such as 'psychologically normal'. (If so, functionalists may treat that predicate as one of the psychological expressions to be defined by Ramsification; so normativity objections in the spirit of Davidson are not an obvious barrier to Ramsified functionalism.) Still others would want $\mathcal{A}$ to include more than just psychology. For example, functionalists convinced of externalism (for example, recent Fodor) could take $\mathcal{A}$ to include an extensive characterization of the subject's external environment and facts about the subject's actual and counterfactual interactions with it; others (for example, Millikan) could take $\mathcal{A}$ to include various facts from the evolutionary history of the subject's ancestors. Functionalists convinced by anti-individualist arguments could take $\mathcal{A}$ to include complex social and linguistic facts pertaining to the subject's community, civilization, or evolutionary lineage. For that matter, $\mathcal{A}$ might well be infinitary (whether law-like or not; Davidson's anomalism is not relevant). Given that our concern is ontological, it makes no difference whether $\mathcal{A}$ is manageable by finite creatures like ourselves. Accordingly, the worry that functionalism is "utopian" (for example, recent Putnam) is not to the point. (Incidentally, it is commonly thought that if functionalists are allowed infinitary sentences $\mathcal{A}$—and, in turn, infinitary Ramsified definitions based on $\mathcal{A}$—then identity theorists should be allowed infinitary disjunctive definitions and that, if they are, the identity theory would be correct and functionalism would have no point. This is quite mistaken, however; see note 33.) Ramsified definitions are extraordinarily flexible and can accommodate an extremely wide range of philosophical views.[8]

[8]But there are limitations on how comprehensive $\mathcal{A}$ may be. If $\mathcal{A}$ includes too much information about the standard mental properties, functionalism is subject to immediate refutation. For example, suppose $\mathcal{A}$ includes a clause stating that, say, the property of being in pain is not a first-order *physical* property (having firing C-fibers, etc.), as is required by the denial of the identity thesis. And so on for the other standard mental properties dealt with by $\mathcal{A}$. In that case, there could be *no* first-order physical properties **R** that satisfy A, contradicting the materialistic version of function-

As indicated, many materialistically inclined functionalists believe that the standard mental properties are not only second-order, but also *physical.* To support their view, they adopt an auxiliary thesis to the effect that, perhaps only as a contingent fact, the first-order properties satisfying A are all physical. If this thesis were correct, these functionalists would then be entitled to say something like this: mental properties are physical, at least as a matter of contingent fact.[9]

1.1 The Argument from Self-Consciousness

Turning now to self-consciousness, consider the following truisms. It is possible for a person to be self-consciously aware that he is in pain. It is possible for a person to be self-consciously aware that he is thinking something. It is possible for a person to be self-consciously aware that he is self-consciously aware of something. It is *not* possible for a person to be self-consciously aware that his mass is greater than mine. It is *not* possible for a person to be self-consciously aware that he has corneas, that he has blue eyes. And so forth. These truisms yield the following principle *I*: a person is self-consciously aware that he is F or that he Fs something only if F is some standard mental property—being in pain, thinking, etc. (that is, one of the properties that functionalists deem to be functionally definable).[10]

---

alism. And more generally, suppose $\mathcal{A}$ includes a clause stating the functionalist thesis that the properties of being in pain, thinking, etc. are not first-order properties. In this case, there could be *no* first-order properties **R** satisfying A, and so the associated Ramsified definitions would be mistaken. The way around these problems is, presumably, to confine $\mathcal{A}$ to a description of the *characteristic interaction* of the standard mental properties with one another and macroscopic physical properties.

[9]In "Some Varieties of Functionalism" Shoemaker gives this view a precise formulation and explicitly advocates it. In their remarks about "physical realizations" of mental properties a great many functionalists seem to be endorsing this materialistic version of functionalism.

[10]This principle is formally consistent with the identity thesis (for example, that being in pain is identical to having firing C-fibers). For example, principle *I* does not formally imply that a person cannot be self-consciously aware that he has firing C-fibers. It only requires that if he is, then having firing C-fibers = being in pain (or thinking or . . .). Of course, principle *I* together with the *negation* of the identity thesis (for example, being in pain $\neq$ having firing C-fibers, etc.) implies that a person cannot be self-consciously aware that he has firing C-fibers. Paul Churchland (in

There are a number of other principles governing self-consciousness. For example, here is a familiar principle that gives a sufficient condition: By engaging in introspection a person comes to be self-consciously aware of the experiences he is having. In a particular case, say, experiencing pain, we have the following principle $\mathcal{P}$: If a person is in pain and engaging in introspection, the person will be self-consciously aware that he is in pain. Of course, qualifiers may be added (for example, '*sharp* pain', '*carefully* engaging in introspection', '*ceteris paribus*'), and $\mathcal{P}$ may be reconstrued as a subjunctive conditional or a conditional probability statement. What matters is that something like $\mathcal{P}$ holds; surely it does. In the ensuing argument I am going to focus on $\mathcal{P}$ for illustrative purposes. I should emphasize that it is just one of many principles describing the characteristic interaction of self-consciousness with other standard mental properties. At least some of those principles must be part of a satisfactory psychological theory $\mathcal{A}$, and that is enough to permit us to construct a parallel argument for the same conclu-

---

"Reduction, Qualia, and the Direct Introspection of Brain States," *Journal of Philosophy* 82 (1985): 8–28) holds that people with prior exposure to physiological theory and its terminology could be self-consciously aware of their brain states. I find this claim incredible. But for the purpose of this paper we may simply sidestep the issue by restricting principle *I* and our applications of it to people who have had no prior exposure to physiological theory and its terminology.

Incidentally, someone might claim that the contents of a being's self-conscious awareness are propositions that involve only species-specific mental properties. But this is wholly implausible. If a being in another species is self-consciously aware that it is in pain, this proposition (that it is in pain) is a predicative proposition formed from the property of being in pain. Likewise, when I am self-consciously aware that I am in pain, this proposition (that I am in pain) is a predicative proposition formed from the property of being in pain. To see that they are one and the same property, note that both propositions have as a *logical consequence* the existential proposition that something is in pain. This proposition—that something is in pain—is something the being and I would *agree* on. This and many other crucial logical relations would be lost if the analysis were based on species-specific properties. (For more on this, see responses 1 and 2 below.) In any case, the appeal to species-specific properties here would violate the spirit of ontological functionalism: if the mental lives of creatures in diverse species are functionally identical, ontological functionalists are committed to holding that they are identical *simpliciter*. Moreover, much the same argument as I am about to give could be mounted against ontological functionalism even if one supposed that the contents of self-conscious awareness were always propositions involving species-specific mental properties.

sions (see responses 4 and 5 below). The underlying problem is unavoidable.

We have agreed that $\mathcal{A}$ contains at least some clauses like $\mathcal{P}$ (perhaps with qualifications). We may suppose without loss of generality that $\mathcal{A}$ is a conjunction of some complex clause $\mathcal{B}$ and $\mathcal{P}$. Let B result from $\mathcal{B}$ by replacing 'is in pain' with the one-place predicate variable 'R_1', 'thinks' with the two-place predicate variable 'R_2', and so on. For simplicity, let us assume that 'introspects' and 'is self-consciously aware' are, respectively, the third and fourth psychological predicates occurring in $\mathcal{A}$. In this case, the functional definitions of introspection and self-consciousness awareness would be fully analogous to the functional definitions of being in pain and thinking. For example, the functional definition of self-conscious awareness would be along the following lines:

> 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 R_1 and R_3, then x will be related by R_4 to the proposition that he is R_1; (iii) x is related by R_4 to the proposition that P.

(Note that clause (ii) results from $\mathcal{P}$ by replacing 'is in pain' with 'R_1', 'introspects' with 'R_3', and 'is self-consciously aware' with 'R_4'.)[11] Let us assume that this definition (or something like it) is correct and that the other functional definitions (for example, of being in pain, introspection, etc.) are correct as well. I will show that the standard mental properties must then be first-order.

Suppose that x is simultaneously in pain and engaging in introspection. Then x would satisfy functional definitions of being in

[11] I am making the following standard supposition, which holds in all classical intensional logics (Frege's, Russell's, Carnap's, Church's, Kaplan's, Montague's, etc.): in principle $\mathcal{P}$ the proposition *denoted* by the 'that'-clause 'that he is in pain' is the same as the proposition *expressed* by the unembedded antecedent 'he is in pain'. In each case, the proposition is formed from that which is expressed by the predicate (verb phrase) 'is in pain'. In their Ramsification of $\mathcal{P}$, our functionalists should therefore use one and the same predicate variable 'R_1' to replace both the embedded and the unembedded occurrence of 'is in pain'. (For further discussion of this point, see response 2 below.) If, instead, embedded occurrences of psychological predicates were not replaced by predicate variables, circularity would result. Alternatively, if principles like $\mathcal{P}$ were just deleted from $\mathcal{A}$, the resulting Ramsified definitions would be too weak. (See response 4 below.)

pain and introspection. Because x is simultaneously in pain and engaging in introspection, the "realizations" **R** that are invoked in the definition of being in pain would be the same as those invoked in the definition of introspection. So by combining the right-hand sides of the two definitions, we may conclude:

> There are first-order properties **R** such that (i) they satisfy B, (ii) if x is R_1 and R_3, then x will be related by R_4 to the proposition that he has R_1, and (iii) x is R_1 and R_3.

Clause (iii) is the antecedent of (ii). So, by modus ponens, (ii) and (iii) imply the consequent of (ii): x is related by R_4 to the proposition that he has R_1. Thus, we have:

> There are first-order properties **R** such that (i) they satisfy B, (ii) if x is R_1 and R_3, then x will be related by R_4 to the proposition that he is R_1, and (iii) x is related by R_4 to the proposition that he is R_1.

This implies (by existential specification and existential generalization on 'R_1') that there is a first-order property F such that:

> There are first-order properties **R** such that (i) they satisfy B, (ii) if x is R_1 and R_3, then x will be related by R_4 to the proposition that he is R_1, and (iii) x is related by R_4 to the proposition that he is F.

But this is the right-hand side of the definition of self-conscious awareness, where the proposition that he is F is put in for the proposition that P. So, given this definition, we may infer the left-hand side: x is self-consciously aware that he is F, where F is a *first-order* property. But principle *I* tells us that x is self-consciously aware that he is F only if F is one of the standard mental properties. It follows that there is a standard mental property (namely, F) that is first-order, contradicting ontological functionalism. Moreover, in the circumstance, the only standard mental property that F could be is the property of being in pain. Thus, the property of being in pain is a first-order property.

This argument easily generalizes to each of the other standard conscious mental properties: all are first-order. And if all these are

first-order, it would be odd in the extreme if the standard nonconscious (for example, dispositional) mental properties were not likewise first-order.

Given that the standard mental properties are first-order, ontological functionalism is mistaken. For, according to it, the standard mental properties are not first-order but rather second-order. Besides ontological functionalism, there are only two credible positions on the ontological status of mental properties: the identity thesis and the thesis that they are first-order nonphysical properties. But there are good reasons to think that the identity thesis is false. (On this point our ontological functionalists—even those who are materialistically inclined—are in agreement.) Given this, we are left with the conclusion that the standard mental properties are first-order nonphysical properties.

This conclusion implies, in turn, that materialistically inclined ontological functionalists are doubly mistaken: standard mental properties are neither second-order nor physical. The conclusion also implies that the materialists' auxiliary thesis (that, at least as a matter of contingent fact, the first-order properties satisfying A are all physical) is false. The reason is that the standard mental properties are first-order satisfiers of A that are not physical.

1.2 Responses

I will now consider a series of responses to the above argument, each of which has been suggested to me by one or more defender of functionalism.

(1) *The Armstrong-Lewis Theory.* Ramsifying functionalists might abandon standard ontological functionalism, according to which mental properties are second-order, and turn to the Armstrong-Lewis "functionalist identity theory," according to which they are first-order physiological properties. On the Armstrong-Lewis theory, an expression like 'pain' is a non-rigid designator for the occupant of the pain-role, which, as a contingent fact, is a first-order physiological property (for example, involving C-fibers). I believe that there are already convincing arguments in the literature against this treatment of 'pain' and kindred expressions.[12] But the

[12]For example, Sydney Shoemaker's "Some Varieties of Functionalism." See also my "Mental Properties."

argument from self-consciousness provides a way to give a general disproof of the Armstrong-Lewis theory.

The point turns on the standard mental properties *being in pain, believing p, desiring q,* etc. As a preliminary, recall the nearly universally accepted principle in philosophy of language that the gerunds ⌜being F⌝ and ⌜F-ing⌝ are rigid designators.[13] For example, 'being the tallest man' denotes the same property in every possible world, namely, the property of being the tallest man. Of course, relative to different possible worlds, different things have that property; for this reason, the *description* 'the tallest man' denotes different things in different worlds. But the *gerund* always denotes the same thing. Likewise for gerunds such as 'being in pain'; they too are rigid designators.

For a reductio, suppose with Armstrong and Lewis that pain $=_{def}$ the first element R_1 in the sequence of first-order **R** satisfying A. Then, to deal with the gerundive form, Armstrong and Lewis would need to accept (something like) the following: being in pain $=_{def}$ being something x such that x has the first element R_1 in the sequence of first-order **R** satisfying A. But this cannot be right. First, the property of being in pain would have to be *second-order,* as standard ontological functionalists believe. Likewise for thinking, introspecting, being self-consciously aware of, etc. But then our original argument may be used, as before, to show that if the

---

[13]Here is a proof of this principle: Predicates (that is, verbs) in a given language do not shift their meaning (versus extension) from world to world—relative to any possible world, ⌜F⌝ means what it means in the actual world. At the same time, relative to any possible world, the gerund ⌜being F⌝, which is the nominalization of the predicate ⌜F⌝, denotes the property expressed by the predicate ⌜F⌝. This is why ⌜x is F iff being F is a property of x⌝ is necessarily true for ordinary ⌜F⌝. It follows that, relative to any possible world, ⌜being F⌝ denotes what it denotes in the actual world. Hence, ⌜being F⌝ is rigid.

On the Armstrong-Lewis theory, the following relativized statements are supposed to hold: in human beings, pain = firing C-fibers; in Martians, pain = firing D-fibers; etc. But the analogous thing certainly does not hold for our canonical gerundive idiom for referring to properties: in human beings, the property of being in pain = the property of having firing C-fibers; in Martians, the property of being in pain = the property of having firing D-fibers; etc. This is wholly counterintuitive, and the above argument shows why. (As shown in note 10, when someone is self-consciously aware that he is in pain, the content—that he is in pain—is a predicative proposition formed from the property of being in pain. And this is so regardless of species.)

Ramsified definitions are correct, all these properties must be *first-order.* Hence, a contradiction. (Suppressing order restrictions does no good; see response 6.) Second, the property of being in pain would have to be identical to the first-order property F identified in our original argument. But on the Armstrong-Lewis theory, F would be a specific physiological "realization" (for example, involving C-fibers). Hence, 'being in pain' would *rigidly* denote some specific first-order physiological "realization," whereas 'pain' would denote such a "realization" only *non-rigidly.* But this is absurd: it implies that there are possible situations in which 'pain' denotes entirely different "realizations" (for example, involving D-fibers) but in which 'being in pain' denotes, as always, the original "realization" (involving C-fibers).

(2) *Nonstandard Intensional Logics.* As indicated (note 11), our argument is based on the following supposition, which holds in all standard intensional logics (Frege's, Russell's, Carnap's, Church's, Kaplan's, Montague's, Lewis's, etc.): In principle $\mathcal{P}$ the proposition denoted by the 'that'-clause 'that he is in pain' is the same as the proposition expressed by the unembedded antecedent 'he is in pain'. In each case, the proposition is formed from that which is expressed by the predicate (verb phrase) 'is in pain'. Indeed, all standard intensional logics embrace the general principle that an atomic sentence $\ulcorner Fa \urcorner$ expresses the proposition denoted by the 'that'-clause $\ulcorner$that Fa$\urcorner$; in each case, the proposition is formed from that which is expressed by the predicate $\ulcorner F \urcorner$, namely, being F. In their Ramsification of $\mathcal{P}$, therefore, functionalists should use one and the same predicate variable 'R_1' to replace both the embedded and the unembedded occurrence of the predicate 'is in pain'.[14]

In an effort to avoid our main argument, some functionalists might challenge standard intensional logic and propose to put in

[14]This is exactly what Russellians would do. Fregeans would do something logically equivalent. They would, for example, rewrite $\mathcal{P}$ with a 'that'-clause in the antecedent: if it is true *that* a person is in pain and engaging in introspection, the person will be self-consciously aware *that* he is in pain. Then they would replace both occurrences of 'is in pain' with one and the same predicate variable 'R_1'. In fact, functionalists who are Fregeans would want to deal with *all* unembedded occurrences of predicates in $\mathcal{A}$ in some such manner. The reason is that they want their predicate variables $\ulcorner R_i \urcorner$ to range over intensions (that is, the sort of entities *expressed* by predicates $\ulcorner F \urcorner$), not over extensions (that is, the sort of entities predicates $\ulcorner F \urcorner$ *refer* to, namely, the *set* of things $\ulcorner F \urcorner$ is true of).

its place a nonstandard intensional logic according to which the proposition denoted by the 'that'-clause 'that he is in pain' differs from the proposition expressed by the unembedded sentence 'he is in pain': whereas the latter is a predicative proposition formed from the property of being in pain, the former is a predicative proposition formed instead from a *concept* of the property of being in pain. There are, however, persuasive arguments against views of this sort.[15] But rather than debating the point, suffice it so say that our argument can be reworked within the envisaged nonstandard logical framework. To see how this would go, note that on this alternate approach $\mathcal{P}$ would be treated along the following lines: if x has the property of being in pain and is engaging in introspection, then x will stand in the relation of self-conscious awareness to a predicative proposition formed from the-concept-of-being-the-property-of-being-in-pain, which is a concept of the property of being in pain. The result of Ramsifying is then: If x has property R_1 and property R_3, then x will stand in relation R_4 to a predicative proposition formed from R_1', where R_1' is a concept of R_1. The rest of the argument would then go through *mutatis mutandis*. Specifically, we can show that R_1' is a concept of a first-order property (for example, having firing C-fibers) and also a concept of the property of being in pain itself. It follows that the property of being in pain is a first-order property, contradicting the ontological functionalists' thesis that it is a second-order property. The point is that somewhere in $\mathcal{A}$ the relation between the-concept-of-being-the-property-of-being-in-pain and the property of being in pain needs to be fixed, so that in the Ramsification of A the relation between the concept R_1' and the property R_1 would correspondingly be fixed, and this allows our argument to go through.[16]

---

[15]See, for example, Jon Barwise and John Perry, "Semantic Innocence and Uncompromising Situations," *Midwest Studies in Philosophy* 6 (1981): 387–404; Bealer, *Quality and Concept* (Oxford: Oxford University Press, Clarendon Press, 1982); Bealer and Uwe Mönnich, "Property Theories," *Handbook of Philosophical Logic* 4 (1989): 133–257. See also many of the authors mentioned in note 30.

[16]Incidentally, someone might try to hold the following: When ⌜F⌝ is a nonpsychological predicate, the proposition denoted by the 'that'-clause ⌜that x is F⌝ is formed from a concept of the property of being F; but when 'F' is a psychological predicate, the proposition denoted by the 'that'-clause ⌜that x is F⌝ is formed, not from a concept of the property of being F, but rather from a concept of a "realization" of the property of being F. This

(3) *The Two-step Approach.* Some functionalists might think that our argument can be avoided by a familiar two-step approach. On this approach, one first attempts to give functional definitions of mental state-types; that is, one attempts to define what it is for a state to be a thought, a desire, etc. Following that, one attempts to define what it is for a mental state of a given type to have p as its content. Putting the two steps together, one then obtains fully general definitions of what it is for x to think p, desire q, etc. For example, x thinks p iff_{def} x is in a state which is a thought and that thought has p as its content.

The two-step approach may be developed by one of two methods. The first relies on Ramsification. For example, state t would be a thought iff_{def} there exist first-order **R** satisfying A and t = the state of being related by R_2 to something. And a state of thought t would have p as its content iff_{def} there exist first-order **R** satisfying A and t = the state of being related by R_2 to p. Plainly, such Ramsified formulations of the two-step approach are straightforwardly subject to our main argument.

The second method incorporates a representationalist theory of mental content built upon causal and/or conceptual roles. I will criticize this formulation in section 2: the problem will be that of defining the content-of relation for mental states whose contents concern mental states themselves. The problem recurs in any two-step theory that aims to characterize the content-of relation without the aid of Ramsification. At the close of section 2, I discuss formulations that combine both methods. I argue that the resulting two-step definitions would be correct only if simpler one-step definitions were correct and that these one-step definitions imply that mental properties are first-order. If these arguments are correct, the two-step approach is a gratuitous complication.[17]

---

logical theory, however, is so disunified and unmotivated that it may not be taken seriously.

[17]In *Inquiry* (Cambridge: MIT Press, 1984), Robert Stalnaker seems to advocate a two-step picture, but one that is designed to do without mental representations. Our main problem threatens both steps. Concerning step one, although Stalnaker does not tell us how he would actually define mental state-types (belief, desire, self-consciousness, etc.), it seems inevitable (given that he does not avail himself of mental representations) that he would need to resort to some sort of Ramsification to avoid standard circularity problems (for example, defining belief in terms of desire and desire in terms of belief). But if he does, it seems that the arguments in the text would carry over *mutatis mutandis*. Concerning step two, Stalnaker

(4) *Alternative Treatments of* $\mathcal{P}$. The next type of response is to revise the way principle $\mathcal{P}$ is treated in Ramsified definitions. This simplest proposal is this: When A is formed from theory $\mathcal{A}$, leave untouched the occurrences in $\mathcal{P}$ of 'is in pain', 'introspects', and 'is self-consciously aware'; do not replace them with predicate variables 'R_1', 'R_3', and 'R_4', respectively. The problem with this proposal is that the resulting functional definitions are circular.[18]

A related proposal would involve simply excluding $\mathcal{P}$ from the theory $\mathcal{A}$ on which the Ramsified definitions are based, that is, by deleting clause (ii) from the definitions. (Likewise for kindred principles that might play $\mathcal{P}$'s role in $\mathcal{A}$.) But this proposal is ad hoc: the original motivation for Ramsified functional definitions was the doctrine that the standard mental properties are defined by the way they behave within a psychological theory describing the characteristic interaction of the standard mental properties with one another—not some artificially weakened theory which leaves out characteristic interactions like those recorded by $\mathcal{P}$. In any case, this functionalist doctrine implies that the weakened definition is no better off than the original. For, according to this doctrine, there are first-order properties **R** that function with re-

---

advocates an ideal-rational-agent version of the reliable-indicator analysis. This analysis has two unacceptable consequences: (1) Believing a proposition entails believing every necessarily equivalent proposition. (2) Believing a proposition entails that one believes that one believes the proposition (and that one believes that one believes that one believes the proposition, *ad infinitum*).

[18]Much the same problem would arise if one proceeded in stages: first, giving functional definitions of all mental properties besides self-consciousness and, second, using these defined notions in a functional definition of self-consciousness. The problem is that there are principles akin to $\mathcal{P}$ in which 'is self-consciously aware' is embedded within itself, for example, a principle whose consequent is this: x is self-consciously aware that he is self-consciously aware of something. How would one Ramsify this clause? If one wrote 'x is related by R_4 to the proposition that he is related by R_4 to something', our main argument could be repeated to show that the relation of self-conscious awareness is a first-order relation, thereby contradicting ontological functionalism. If instead one left the consequent of the relevant clause untouched (that is, 'x is self-consciously aware that he is self-consciously aware of something') or if one wrote 'x is related by R_4 to the proposition that he is self-consciously aware of something', the definition would be viciously circular. Of course, one might try to avoid the problem by just deleting the relevant principle from the theory $\mathcal{A}$, but this proposal would be subject to the problem I am about to discuss in the text.

spect to each other in the characteristic way that the standard mental properties do. But the characteristic way the standard mental properties function with respect to each other is given by $\mathcal{A}$ (that is, $\mathcal{B}$ & $\mathcal{P}$). Accordingly, they satisfy A (that is, B & P) and, in turn, B alone. It follows that if the functional definition based on B were correct, a person could be self-consciously aware that he has R_1, where this is a first-order property. But the only property R_1 could be is the property of being in pain. So being in pain would be a first-order property, contrary to ontological functionalism.

Motivating these reformulations is the idea that principle $\mathcal{P}$ is the source of the difficulty and that the difficulty can therefore be avoided by somehow modifying $\mathcal{P}$'s role in functional definitions. But this fails to appreciate the real problem. Principle $\mathcal{P}$ is symptomatic of a general problem. In a satisfactory psychological theory describing the characteristic interaction of self-consciousness with other standard mental properties, there would be a variety of principles that contain standard psychological predicates embedded within the scope of 'is self-consciously aware that' and that place corresponding requirements on the range of the relation of self-consciousness awareness.[19] Focusing on these other principles, we can construct analogues of the above argument. What is needed is a *systematic* response to the problem of multiply embedded psychological predicates.

One systematic response would be to suppress *all* such principles in the Ramsified definition of self-consciousness. But this response is plainly unsatisfactory: there would be nothing in the resulting definition to restrict the range of R_4, and so it would let in too much.

(5) *Psychofunctionalism.* Another systematic response would be to try to skirt self-consciousness altogether by holding that although it plays a role in "folk" psychology, it is eliminable from a "scientific" psychology.[20] The resulting "scientific" psychology $\mathcal{A}$

[19]Some of these principles are like $\mathcal{P}$ in that they specify "dynamic" characteristics of standard mental properties. Others are "static," simply restricting the range of the relation of self-consciousness awareness. *I* itself is one such principle: for all *F*, if a person is self-consciously aware that he is *F* or that he *F*s something, then either *F* = being in pain or *F* = thinking or .... Unlike $\mathcal{P}$, *I* gives a necessary not sufficient condition for self-consciousness.

[20]Personally, I find this idea wholly implausible: scarcely anything could ever be better justified to someone like you or me than the fact that he is

would treat a truncated list of mental properties and would serve as the basis for their Ramsified definitions.

This proposal, however, encounters much the same sort of problem as the previous one. The reason is that even if a "scientific" psychology were oriented exclusively toward the mere explanation of behavior, functionalists would still recognize the need for principles that concern embedded psychological states and that place substantive requirements on the ranges of associated psychological relations. For example, there would be principles concerning "self-monitoring" on the part of intelligent beings. The following is an illustration of such a principle: If someone were entertaining the question of whether he is in pain and if he were indeed in pain, he would think that he is in pain. (As before, qualifiers—for example, '*sharp* pain', '*seriously* entertaining', '*ceteris paribus*'—may be added, and the principle may be reconstrued as a conditional probability statement.) Such a principle might play a role in explaining why people respond as they do to queries about their current mental states. In addition to this principle, there would be principles (perhaps qualified) about "intrinsic" desires or preferences regarding one's own psychological states (for example, *ceteris paribus* desire for pleasure, happiness, knowledge, love, etc.; *ceteris paribus* aversion to pain, nausea, fear, etc.). Perhaps there would also be principles governing the states of "mutual knowledge" required for cooperative social activities (especially linguistic activity). The point is that even a "scientific" psychology would have to contain a plethora of principles concerning embedded psychological states. Since such principles describe the characteristic behavior of the associated psychological relations and place substantive requirements on them, they would need to be included in the theory $\mathcal{A}$ upon which our functionalists would base their Ramsified definitions. But given this, we could mount the same sort of argument as before.

So our functionalists are still in need of a systematic response to the problem of multiply embedded psychological predicates. The only other candidate I know of would be the adoption of a thor-

---

self-consciously aware of various things; it would be altogether unscientific of him to omit principles concerning this undeniable phenomenon. Moreover, in normal cases, when a person knows that he is self-consciously aware, say, that he is thinking, this is neither a *theory* in any standard sense of the term, nor "theory-laden."

oughgoing "representationalism" according to which embedded occurrences of mental predicates are not used in their ordinary way but instead are used to *mention* "mental representations." If the latter were *linguistic expressions,* then standard use/mention considerations would show that associated embedded occurrences of mental predicates should not be replaced by predicate variables when we Ramsify $\mathcal{A}$. In particular, we should not replace the mental predicates embedded within the scope of 'is self-consciously aware'; accordingly, our argument would be blocked. Of course, if embedded occurrences of mental predicates were only used to mention expressions, the same ought to hold for occurrences of other embedded expressions. The upshot would be that all occurrences of expressions within the scope of mental predicates would merely mention expressions. This would suggest the view that mental states consist of subjects standing in relations to sentences (that is, to concatenations of these mentioned expressions). Those sentences surely would not belong to any natural language (there are good arguments against that idea), but they might belong to some hypothetical language such as Wilfrid Sellars's "Mentalese." The resulting picture would be a version of the second main form of functionalism, namely, language-of-thought functionalism. In section 2, I will argue that language-of-thought functionalism does not avoid our problem but only puts it off. Before I do this, I should mention a final attempt to save Ramsified functionalism.

(6) *Altering the Order Restrictions.* A final response to our main argument is to modify the order restrictions within the Ramsified definitions. One way to do this is to reformulate the Ramsified definitions so that **R** is restricted to *second-order* properties rather than first-order properties. For example,

> x is self-consciously aware that P iff_{def} there exist second-order properties **R** satisfying A such that x R_4 P.

The problem is that self-conscious awareness would have to be *third-order* rather than second-order. But then, by applying our argument again, we can show that if this definition were correct, self-conscious awareness would have to be second-order. Again, a contradiction.

Whatever order restriction is imposed on **R**, the same sort of

problem recurs. This suggests that functionalists might try to avoid the problem by deleting *all* order-restrictions on **R**.[21] For example,

> x is self-consciously aware that P iff_{def} there exist properties **R** satisfying A and x is related by R_4 to the proposition that P.

The idea is to permit properties of any order to be in **R.** For example, the property expressed by the *entire* right-hand side is permitted as one of the values of the predicate variable 'R_4' that occurs *within* the right-hand side.

When one drops the earlier, restricted, definition in favor of this new, unrestricted, definition of self-conscious awareness, our argument cannot be given in its original form. But there is a straightforward variation which shows that if this unrestricted definition is correct, the *only* properties that can ever satisfy A are the standard mental properties themselves. (Roughly, if there were satisfiers of A other than the standard mental properties, the definition would wrongly admit propositions involving them into the range of self-consciousness, thereby contradicting principle *I*.) That is, properties distinct from the standard mental properties cannot satisfy A, and so there are no distinct, ontologically prior properties in terms of whose pattern of A-like interaction the standard mental properties can be defined. In an *ontological* sense, therefore, the standard mental properties are first-order: the properties whose existence is asserted in the envisaged definitions are the very properties being defined. Unlike properties that are ontologically second-order, they are not defined by the general pattern of causal (or functional) interaction of ontologically prior properties. Thus, our conclusion stands.

If the envisaged unrestricted definitions should happen to be correct, the standard mental properties could be called 'second-order' only in the mere *linguistic* sense that they are definable with the use of predicate variables. This is not the sense of 'second-

---

[21]There is an explicitly two-tier proposal that also omits order restrictions: for example, x is self-consciously aware that P iff_{def} there exist properties **r** and properties **R** both satisfying A such that **r** are "realizations" of **R** and x R_4 P. This proposal is defeated by problems analogous to those cited at the end of section 2 and in note 45. There is also a more complicated two-tier proposal that weaves **R** into the propositions comprising the ranges of **r**, but it too is defeated by such problems.

order' intended by ontological functionalists: they use it to state a substantive thesis on the ontological status of mental properties. In Quine's terminology, the unrestricted definitions, if correct, would have only "ideological" significance. Perhaps there are people calling themselves functionalists who have only ideological ambitions, that is, who aim only to identify the standard mental properties using exclusively nonmental vocabulary and whose only goal is conceptual clarification. (In section 3, I will discuss whether ideological functionalism can achieve this far more modest goal.) To get a better feel for the distinction between ideological functionalism and ontological functionalism, consider an analogy. Suppose that phenomenal colors are ontological primitives and, hence, ontologically first-order. Then Hume's missing shade of blue—call it 'humeblue'—has a Ramsified definition that is second-order in the mere linguistic sense: x is humeblue iff$_{def}$ there is a property F such that F is a shade of blue and F $\neq$ blue$_1$ and F $\neq$ blue$_2$ and . . . and x has F. Plainly, this "ideological" definition of humeblue does not show that humeblue is ontologically second-order. For much the same reason, "ideological" functionalism is consistent with our conclusion that, ontologically, the standard mental properties are first-order properties.

This brings out a general point. For any definable property, there will always be "definitions" that correctly pick out the property but that are, linguistically, of arbitrarily high order; but these linguistically higher-order definitions have no ontological significance. This will be relevant at the close of section 2, where I will compare a simple unrestricted Ramsified definition like that under discussion here and a more complicated Ramsified definition that needlessly incorporates a layer of mental representations. We will see that the latter definition correctly picks out the intended mental property only if the former does as well. By the argument indicated two paragraphs above, however, we know that the simple unrestricted definition would be correct only if the property were ontologically first-order. So even if the more complicated definition happened to pick out the property correctly, the property would be ontologically first-order.

Let us sum up. In our survey of responses, we have found only one way that ontological functionalists might plausibly try to avoid our conclusion that mental properties are first-order, namely, by retreating to the sort of thoroughgoing representationalist theory

of mental content described above. Language-of-thought functionalism typifies this view. According to it, mental states consist of subjects standing in relations to sentences in a hypothetical language of thought. I will now argue that this theory runs into much the same problem we have been discussing, only at one step removed.

2. Language-of-Thought Functionalism

In most of my discussion I will employ Schiffer's convenient "token-in-a-box" metaphor; my comments apply *mutatis mutandis* to non-metaphorical formulations stated in terms of abstract relations to sentences in a language of thought. In the token-in-a-box idiom, mental states are identified with tokenings of Mentalese sentences in "modules"—a Belief Box, a Desire Box, etc. According to the theory, x believes p iff_{def} there is a Mentalese sentence s tokened in x's Belief Box and s's content is p; x desires p iff_{def} there is a Mentalese sentence s tokened in x's Desire Box and s's content is p; and so forth.[22] Note that these analyses are not complete until the terms on the right-hand sides—notably, the relation *being the content of*—are defined, and they are not satisfactory unless this can be done without circularity.

Language-of-thought functionalists take the content of a com-

---

[22]See, for example, Jerry A. Fodor, *Psychosemantics* (Cambridge: MIT Press, 1987). I believe that token-in-a-box functionalism cannot literally be right, for it is inconsistent with the possibility of disembodied subjects—a possibility that many functionalists accept and for which supporting arguments can be given (see my "Mental Properties"). If the token-in-a-box theory were taken literally, the "boxes" and the "tokens" would have to be functioning bodily parts. Given that a disembodied subject would have no functioning bodily parts, a literal formulation of language-of-thought functionalism must therefore be stated, not in terms of "tokens" and "boxes," but rather in terms of abstract relations to language-of-thought sentences (versus sentence tokens). An alternative response to the possibility of disembodiment would be to offer token-in-a-box functionalism, not as a *general* theory of mind, but rather as a theory of the *human* mind, assuming that human beings are essentially embodied. But this restricted token-in-a-box theory would falter over the problem I am about to discuss in the text, namely, the problem of specifying the contents of (human) Mentalese psychological predicates. The reason is that human thoughts about mental properties are not typically about human-thinking, human-desiring, human-self-consciousness, etc.; rather they are about thinking, desiring, self-consciousness, etc. *simpliciter.* (See note 10.)

plex Mentalese expression to be compositional: they assume that the expression's content is determined by its form plus the contents of its constituent primitive Mentalese expressions. They hope to define the content-of relation for primitive Mentalese expressions in functional or causal terms. There are two basic ways of doing this: "conceptual roles" or "world-word" causal relationships. Of course, the two approaches could be synthesized into a single "two-factor" analysis of content.[23] (In the ensuing discussion I will consider simplified versions of the two approaches but will be careful not to let the simplifications prejudice the discussion; the problems I will uncover have an altogether different source.)

The following examples illustrate how a conceptual role analysis might go. Suppose that a category of Mentalese logical constants (say, '∃', '&', '¬', etc.) has been identified. Then, if it is nomologically necessary that '∃' behaves in Mentalese sentences in the Belief and Desire Boxes in the way 'there exists' behaves in good arguments in rational-decision theory, the content of '∃' $=_{\text{def}}$ the operation of existential generalization.[24] Analogously for the other

---

[23]For an example of a conceptual role approach, see Gilbert Harman, "Conceptual Role Semantics," *Notre Dame Journal of Formal Logic* 23 (1982): 242–56. For a world-word causal approach, see Dennis Stampe, "Toward a Causal Theory of Linguistic Representation," *Midwest Studies in Philosophy* 2 (1977): 42–63; and Fred Dretske, *Knowledge and the Flow of Information* (Cambridge: MIT Press, 1981). For a sophisticated descendant of the world-word causal analysis, see Jerry Fodor, *Psychosemantics*; my discussion of the causal theory will apply *mutatis mutandis* to this view. For an example of a "two- factor" approach, see Ned Block, "Advertisement for a Semantics for Psychology," *Midwest Studies in Philosophy* 10 (1986): 615–78.

[24]Strictly speaking, the definition of the content-of relation for Mentalese primitives would be a general "definition-by-cases" in which the definition just given in the text is associated in the obvious ways with one of the cases: x is the content of y iff_{def} (1) if y is a logical constant such that it is nomologically necessary that y behaves in Mentalese sentences in the Belief and Desire Boxes in the way 'there exists' behaves in good arguments in rational-decision theory, then x = the operation of existential generalization and (2) . . . .

Incidentally, conceptual role functionalists need to be careful in how they characterize the notion of a second-order property. Their notion should be relative, not absolute. Specifically, they should relativize (or parameterize) their notion to an antecedently given class of nonpsychological entities, among which may be nonpsychological entities of arbitrary order. For example, suppose existential generalization is a second-order intension. Would this force conceptual role functionalists to hold that mental properties are of some order higher than two? No. For, on the charitable

logical constants. Here is another sort of example: Mentalese would have a counterpart of the first-person singular pronoun so that a subject can have first-person beliefs about itself. The conceptual role theory can neatly identify this expression—suppose it is 'i'—by exploiting the characteristic pattern of its tokenings in the subject's respective Boxes. They would then define its content thus: the content of 'i' $=_{def}$ the subject itself. (Notice that a plurality of Boxes is implicated in each of these treatments.)

According to the world-word causal approach, the content of a Mentalese expression is, roughly speaking, that item (object, property, or relation) in the world that, in the relevant way, causes the expression to be tokened in the Belief Box. (Unlike a conceptual role analysis, this approach focuses exclusively on activity in the Belief Box; the other boxes are idle.) A causal analysis would go along roughly the following lines: Suppose that a category of Mentalese demonstratives (say, 'd_1', 'd_2', 'd_3', etc.) for mid-sized physical objects has been identified and their contents somehow defined. And suppose a category of Mentalese predicates (say, 'F_1', 'F_2', 'F_3', etc.) for mid-sized physical objects has also been identified. Then the content-of relation for such predicates might be defined along (something like) the following lines: the content of F_j $=_{def}$ the property G such that, for each Mentalese demonstrative d_k and mid-sized physical object z that is the content of d_k, z's G-ing would in normal conditions cause $\ulcorner F_j d_k \urcorner$ to be tokened in the subject's Belief Box. It would follow, for example, that the content of 'C' would be the property of being a cow iff, for each Mentalese demonstrative d_k and mid-sized physical object z that is the content of d_k, z's being a cow would in normal conditions cause $\ulcorner Cd_k \urcorner$ to be tokened in the subject's Belief Box.

In my opinion, even for the most elementary cases these two approaches to defining the content-of relation are fraught with problems. But let us suppose that these theories, or variants of

---

reading, they mean that mental properties are of order two, relative to an antecedently given class of nonpsychological entities. This brings up a related point. In Russell's own ramified theory, propositions themselves have orders, and this affects how he would classify relations holding between individuals and such propositions. Functionalists have suppressed this (unnecessary) complication and, at least in this regard, they treat propositions on a par with individuals. This is certainly coherent, and in this paper I am following them on this point.

them, can be made to work for the indicated categories of Mentalese expressions. The question we need to consider is whether these theories succeed in the case of the Mentalese constants for the standard mental relations themselves. For example, according to language-of-thought functionalism, a person believes that he believes something iff the Mentalese sentence '(∃x) i B x' is tokened in his Belief Box. And the Mentalese predicate 'B' should have as its content the belief relation itself (that is, the relation holding between a subject and a proposition such that the subject believes the proposition). Likewise for other multiply embeddable attitudes and the Mentalese psychological predicates ('D', etc.) that are supposed to have them as their contents. Does the causal theory of content yield these results? Does the conceptual role theory? True, at least provisionally, we might have been told what it is for a *state* to be a belief, what it is for a *state* to be a desire, what it is for a *state* to be a self-conscious awareness, etc. But we have not been told what it is for such states to have a specific kind of content, namely, the content expressed by a Mentalese sentence containing psychological relational predicates ('B', 'D', etc.). It is here that the problem that confronted Ramsified functionalism resurfaces.

I will argue that both approaches are doomed. The causal approach, however, is subject to a number of difficulties that do not beset the conceptual role approach. So, at least initially, the latter approach will seem more promising. Let us then begin the discussion with the causal approach.

2.1 The Causal Approach

When the causal approach is extended from macroscopic physical properties to mental relations, we get something like this (remember, a causal analysis focuses exclusively on activity in the Belief Box): the content of 'B' $=_{\text{def}}$ the relation R such that, for each Mentalese sentence s and proposition p that is the content of s, the subject's bearing R to p would in normal conditions cause ⌜i B s⌝ to be tokened in the subject's Belief Box. Similarly, the content of 'D' $=_{\text{def}}$ the relation R such that, for each Mentalese sentence s and proposition p that is the content of s, the subject's bearing R to p would in normal conditions cause ⌜i D s⌝ to be tokened in the subject's Belief Box. Likewise for other Mentalese psychological predicates.

This causal analysis suffers from two main problems, which at bottom have the same source. The first problem is that the analysis is *circular*: the content-of relation for sentences is invoked on the right-hand side; but in order to define this relation, one must first define the content-of relation for the primitive predicates—including in particular 'B', 'D', etc.—that occur in those sentences.[25] In this regard, the definition of the content-of relation for Mentalese psychological predicates ('B', 'D', etc.) is fundamentally worse off than the definition of the content-of relation for Mentalese mid-sized physical-object predicates ('F_1', 'F_2', 'etc.). The latter definition presupposes the content-of relation for the arguments of those predicates, namely, Mentalese demonstratives designating mid-sized physical objects ('d_1', 'd_2', etc.). It is at least plausible that this could be done independently. In the case of Mentalese psychological predicates, however, their arguments include Mentalese sentences that sometimes contain Mentalese psychological predicates themselves. So there is no hope of having an independent definition of the content-of relation for these arguments. Here the definition is in principle circular. (I will develop this point further in my discussion of a kindred circularity problem in the conceptual role approach.) A natural response to this circularity problem is to resort to Ramsification. But then the argument of section 1 can be mounted again, showing that the standard mental relations are first-order and, moreover, that the excursion into language of thought is a gratuitous complication. (See "Retreat to Ramsification" below.)

To see the second problem, notice that the causal analysis could be right only if the relation R (that is, the content of 'B') were the

---

[25]The following alternate analysis avoids the circularity but at the expense of getting the wrong relation: the content of 'B' $=_{def}$ the relation R such that, for each Mentalese sentence s, the subject's bearing R to s would in normal conditions cause ⌜i B s⌝ to be tokened in the subject's Belief Box. After all, the belief relation is not a relation between subjects and Mentalese *sentences*; it is a relation between subjects and the *propositions* that are supposed to be the contents of those Mentalese sentences.

Incidentally, the following analysis also fails: the content of 'B' $=_{def}$ the relation R such that, for *some* Mentalese sentence s and proposition p that is the content of s, the subject's bearing R to p would in normal conditions cause ⌜i B s⌝ to be tokened in the subject's Belief Box. The analysis fails because relations such as considering and entertaining would also satisfy the definition (see note 27 for further explanation).

belief relation itself. (Likewise for analyses of the content of every other Mentalese psychological predicate and the property or relation R that is supposed to be its content.) Recall also that language-of-thought functionalists propose to define the belief relation in terms of the content-of relation: x believes p iff_{def} a Mentalese sentence whose content is p is tokened in x's Belief Box. When this definition is spelled out fully (that is, when the definition of the content-of relation is plugged into the right-hand side), the quantified variable 'R' would occur just as it did above and would have to have the belief relation as its value.[26] The primary tenet of ontological functionalism, however, is that the standard mental properties and relations (for example, belief) can be defined wholly in terms of the general pattern of causal interaction of first-order properties and relations. If this tenet were correct, the indicated occurrence of 'R' would therefore have to have a first-order relation as its value. It would thus follow that the belief relation must be first-order. (Call this the levels problem.) The point is that the belief relation is not being defined wholly in terms of the general pattern of causal interaction of ontologically prior "realizations," as ontological functionalists require. On the contrary, the proposed definition could be correct only if it endowed the belief relation with an ontological primacy incompatible with the ontological functionalists' basic picture. (Plainly, this problem also arises for other Mentalese psychological predicates and the properties and relations that are supposed to be their contents.)

These two problems—the circularity problem and the levels problem—could be avoided if the causal theory were carefully stated within a traditional type-theoretic logical framework. In our discussion of the conceptual role approach, however, we shall see that

[26]The same thing would hold if the belief relation were defined directly as: x believes p iff_{def} there is a unique relation R such that, for each Mentalese sentence s and proposition q that is the content of s, the subject's bearing R to q would in normal conditions cause ⌜i B s⌝ to be tokened in the subject's Belief Box, and x is related by R to p. There would be no objection to this definition (or that in the text) if it could be rewritten as an inductive definition in which the offending occurrences of 'R' could be eliminated and in which the only surviving occurrences of predicate variables had as values first-order "realizations." But the causal theory is incompatible with this idea: the belief relation is not being built up inductively; rather it is picked out all at once as an independently existing, causally efficacious relation.

this framework yields an essentially unsatisfactory treatment of the psychological attitudes and so may not be invoked to save either approach. Accordingly, my final assessment is that the circularity problem and the levels problem are fatal to the causal approach.

Before proceeding to the conceptual role approach, I will briefly describe some further problems with the causal approach. Suppose that the circularity problem and the levels problem did not exist. The majority of functionalists would be still forced to admit that the causal analysis does not succeed. Specifically, given their views on causation, these functionalists would be forced to hold that the causal approach would fail to assign *any* content to Mentalese psychological predicates. Consider the case of 'B'. On the majority view, particular tokenings of the Mentalese sentence s in one's Belief Box are what cause tokenings of ⌜i B s⌝; standing in relation to the *proposition* p (that is, s's content) is causally inert with respect to tokenings of ⌜i B s⌝ in one's Belief Box. (Remember, tokenings in the human case are *physical* tokenings in our brains.) For these functionalists, the relevant causal connections are always at the level of the "realizations": a tokening of a Mentalese psychological predicate is caused by other Mentalese tokenings; mental relations to propositions do not have the power to produce tokenings of Mentalese psychological predicates. As Jerry Fodor says, "[E]ven though it's true that psychological laws generally pick out the mental states that they apply to by specifying the intentional contents of the states, it *doesn't* follow that intentional properties figure in the psychological mechanisms. And while I'm prepared to sign on for counterfactual-supporting intentional generalizations, I balk at intentional causation" (*Psychosemantics,* 140; emphasis in the original). Fodor and functionalists like him require the causation to be at the level of the "realizations," not at the level of the mental relations themselves. On their view, therefore, there could be no relation R having the causal properties required by the definition of the content of 'B', so that definition would assign *no* content to 'B'. (Let us call this the problem of mind-brain causation.)

To avoid this problem, some causal theorists might try to be more liberal about the causal role of mental relations in the production of tokenings of Mentalese sentences. These causal theorists would be willing to hold that standing in a relation R (for example, the belief relation or the desire relation) to proposition p can really cause tokenings in one's Belief Box of Mentalese sentences

such as ⌜i B s⌝ or ⌜i D s⌝. A worry about this position, however, is that it might accord to R an ontological primacy that is incompatible with ontological functionalism. The worry goes as follows: The subject's standing in relation R to p would have to be causally on a par with, say, the presence of a cow in the subject's perceptual field; both must be full-fledged causes right here in the world. And this might imply that the relation R and the property of being a cow must be ontologically on a par. Because the latter is first-order, so is the former. Now causal theorists may try to respond to this worry. (For example, they might hold that in the case of R there is a novel kind of causation at work, namely, one that is *necessarily mediated* by first-order "realizations.") The point I want to make is that the causal approach to content is saddled with a thorny metaphysical problem that simply does not arise on the conceptual role approach. Other things being equal, this a reason to prefer the latter.

There are other reasons to favor the conceptual role approach over the causal approach to Mentalese psychological predicates. Notice that the proposed causal analysis presupposes a "BB-principle"—that is, a principle to the effect that if a person believes p, he believes that he believes p. (This presupposition is manifested in the causal theorists' principle: if ⌜s⌝ is tokened in the subject's Belief Box, so is ⌜i B s⌝.) Similarly, these causal theories presuppose a "BT-principle"—that is, if a person is thinking p, he believes that he is thinking p. And so on for each of the other standard psychological relations. But such principles are false: they are far too strong. For example, BB wrongly implies that anyone who believes p automatically has infinitely many associated beliefs—he believes that he believes p; he believes that he believes that he believes p; *ad infinitum.*

The indicated principles (BB, BT, etc.) would hold only if they were suitably qualified. But these qualifications would need to involve *further* psychological notions. For example, in the case of BT the qualified principle would be something like this: if in normal cognitive conditions a person is thinking p and is considering the question whether he is thinking p, he would believe that he is thinking p. (Let appropriate auxiliary qualifiers be added if needed—'*completely* normal cognitive conditions', '*explicitly* thinking p', '*carefully* considering the question whether he is thinking p', etc.) Notice, though, what is happening. We are invoking a principle

that characterizes the *interaction* of thinking, considering, and believing in normal cognitive conditions.[27] This sort of interaction is precisely the sort of thing that is characteristic of a conceptual role approach. In this way, causal approaches to the content of Mentalese psychological predicates give way to a (form of) conceptual role approach.[28]

2.2 The Conceptual Role Approach

How would a conceptual role theory specify the content of Mentalese psychological predicates—for example, 'T', the Mentalese predicate whose content is supposed to be the relation of thinking? Suppose one mechanically mimicked the style of conceptual role analysis sketched at the outset of this section. The resulting conceptual role analysis would then be something like this: If it is nomologically necessary that 'T' behaves in sentences in the respective Boxes (Thinking Box, Considering Box, Belief Box, etc.) in the way the predicate 'thinks' behaves in psychological theory $\mathcal{A}$, then the content of 'T' $=_{\text{def}}$ the relation of thinking. As it stands, the analysis is viciously circular; the relation of thinking is mentioned on the right-hand side of the definition. But this is one of the relations that, ultimately, we are trying to define. The same holds for each of the other standard mental relations—desiring,

---

[27]Note in this connection that causal theorists would have to deem both the thinking and the considering to be causes of the believing. This is relevant to the point made at the close of note 25.

[28]Suppose that BT were modified with *every* relevant qualifier of the sort just indicated. This new BT principle would not be a mere causal or nomological necessity. It would be necessary *tout court.* (Indeed, if this new BT principle were merely a causal necessity, we would not be able to know it in the way we evidently do, that is, as something which is intuitively obvious in the way many other necessary truths are.) But if the relation between the antecedent and consequent of this BT principle is necessary *tout court* and if the relations between the "realizations" of mental properties really mimic the relations between the corresponding mental properties, the relation between the "realization" of the antecedent of BT and the "realization" of the consequent of BT would also have to be necessary *tout court.* On the usual causal picture, however, this relation ought to be only causally or nomologically necessary. (Call this the modal problem.) Since the stronger modality meshes readily with a conceptual role picture, here is another reason why language-of-thought functionalists might, in the case of Mentalese psychological predicates, be led to abandon causal models and turn to a conceptual role account.

believing, etc. In a traditional type-theoretic framework, however, this circularity problem—as well as all the other problems that confronted the causal approach—would not arise. Let me explain.

According to traditional type theorists, the propositional-attitude verbs 'thinks', 'desires', etc. are "typically ambiguous." There is not a single relation of thinking, a single relation of desiring, etc. Instead there are hierarchies of mental relations (for example, thinking$_0$, thinking$_1$, thinking$_2$, . . . ; desiring$_0$, desiring$_1$, desiring$_2$, . . . ; etc.). And so in Mentalese there would be an associated hierarchy of psychological predicates ('T_0', 'T_1', 'T_2', . . . ; 'D_0', 'D_1', 'D_2', . . . ; etc.). If there were such hierarchies, one would have a promising way of avoiding the above problems. To formulate the theory, one would proceed as follows. First, one would use conceptual roles to identify which Mentalese predicates are predicates for standard mental relations. This would be done as above: the candidates ('T_0', 'T_1', 'T_2', . . . ; 'D_0', 'D_1', 'D_2', . . . ; etc.) are those Mentalese predicates whose behavior in sentences in the respective Boxes (Thinking Box, Belief Box, Desire Box, etc.) matches that of the typed predicates 'thinks$_0$', 'thinks$_1$', 'thinks$_2$', . . . ; 'desires$_0$', 'desires$_1$', 'desires$_2$', . . . ; etc. in a type-theoretical formulation of psychological theory $\mathcal{A}$. Second, using either a causal or conceptual role approach, one would define a content-of relation for all nonpsychological expressions (nonpsychological constants and sentences containing no constants beyond these). Then, for Mentalese psychological predicates, one would offer the following hierarchy of definitions: x thinks$_0$ p iff$_{def}$ some nonpsychological Mentalese sentence whose content is p is tokened in x's Thinking Box. The relation of thinking$_0$ would thereby be defined wholly in terms of the general pattern of behavior of ontologically prior properties and relations, as functionalists require. So one would be entitled to define the content of 'T_0' thus: the content of 'T_0' $=_{def}$ the relation of thinking$_0$. Next, for $n > 0$, one would have the following definition: x thinks$_n$ p iff$_{def}$ some level $n-1$ Mentalese sentence whose content is p is tokened in x's Thinking Box. The relation of thinking$_n$ would thereby be defined in terms of antecedently defined, ontologically prior properties and relations, as functionalists require. Accordingly, one would be free to define the content of $\ulcorner T_n \urcorner$ as follows: the content of $\ulcorner T_n \urcorner =_{def}$ the relation of thinking$_n$.

This approach avoids each of the problems discussed above.

First, it avoids the circularity problem, for at each stage every item on the right-hand side will already have been defined. Second, it avoids the levels problem, for at each stage no property or relation is invoked unless it is nonpsychological or antecedently defined and hence ontologically prior in the sense required by ontological functionalists. Third, it avoids the problem of mind-brain causation. Since the content of 'thinks$_0$' (that is, thinking$_0$) is a derived relation defined in terms of the general pattern of tokenings of Mentalese nonpsychological expressions, the present approach may remain neutral on the question of whether the content of 'thinks$_0$' is a relation having causal powers comparable to standard first-order relations. Analogously, for 'thinks$_1$', 'thinks$_2$', and so on up the hierarchy. Finally, this approach avoids the BB problem because no such (false) principle is included in the psychological theory $\mathcal{A}$; instead, $\mathcal{A}$ contains a correctly qualified counterpart of the principle.[29]

Unfortunately, the logical framework—"typical ambiguity," etc.—upon which this approach is based has been discredited; today most philosophical logicians[30] accept that the psychological relations are in the relevant sense *type-free.* In the background are arguments like those Kripke[31] and others have given to show that 'true' in ordinary language is not typically ambiguous ('true$_0$', 'true$_1$', 'true$_2$', . . .), as Tarskians would have us believe. Useful and intelligible ordinary conversations (such as those between Nixon and Dean, described by Kripke) would make no sense if 'true' were subject to typical ambiguity, having a hierarchy of distinct senses. There are wholly analogous arguments showing that the propositional-attitude verbs are not subject to typical ambiguity. The point can be dramatized by recalling Descartes's *cogito.* When I go through the *cogito,* I think that I am thinking something. (In symbols: i Think [(∃p) i Think p].) The proposition I think—namely,

---

[29]In turn, this approach avoids the modal problem (note 28); for, in connection with the indicated qualified principle, $\mathcal{A}$ does not wrongly attribute a mere causal connection but rather the relevant stronger modality.

[30]The list is long, and includes Aczel, Bealer, Chierchia, Davidson, Dunn, Feferman, Fine, Fitch, Gaifman, Gilmore, Gupta, Jubien, Kripke, McGee, Menzel, Mönnich, Martin and Woodruff, T. Parsons, Perry and Barwise, Reinhardt, Salmon, Soames, and Turner.

[31]"Outline of a Theory of Truth," *Journal of Philosophy* 72 (1975): 690–716.

that I am thinking something—is formed from the relation of thinking, not from any of the hypothesized relations thinking$_0$, thinking$_1$, thinking$_2$, etc. Nor is the relation of thinking the union of these hypothesized relations. Indeed, it is possible for me to be thinking that I am thinking something and, for each $n \geq 0$, to be thinking$_n$ nothing whatsoever. That is, I could be thinking that I am thinking something and, at the same time, not be thinking$_0$ any proposition (that is, any nonpsychological proposition), and not be thinking$_1$ any proposition (that is, any proposition about thinking$_0$, desiring$_0$, etc.), and so forth.[32] Likewise, I could be thinking that I am thinking$_n$ something, but not be thinking that I am thinking something. Reflections like these show, not only that the standard propositional attitudes are type-free, but also that they are *ungrounded*: We can have attitudes toward type-free general propositions that are about nothing other than those very attitudes. When I go through the *cogito,* the proposition I am thinking—namely, that I am thinking something—is true simply because I am thinking that very proposition.[33] Thought floats freely upon itself.

---

[32]If you doubt this, shift to the following variant of the *cogito*: I am contemplating more intently than anything else the proposition that there is something I am contemplating more intently than anything else. In symbols: 'i C [(∃p) i C p]'. I can be contemplating this proposition more intently than anything else and yet, for each $n \geq 0$, I might be contemplating$_n$ nothing whatsoever (that is, contemplating$_0$ no nonpsychological proposition, contemplating$_1$ no proposition about contemplating$_0$, and so forth).

The points in the text show that the following fails: x thinks p iff for some n, x thinks$_n$ p. Of course, *within* standard type theories one cannot even quantify over levels n.

[33]The propositional attitudes thus differ from truth (as Kripke explains the concept in "Outline"): once the extension of 'true' over the nonsemantical is fixed, its extension over the semantical is fixed as well. The extension of 'thinks' is formally different: even once the extension of 'thinks' over the nonpsychological is fixed, its extension over the psychological remains open.

In this connection we can see that infinitary disjunctive definitions of mental relations fail because they are circular: for example, x thinks p iff$_{def}$ (x is in physical state S and p = the proposition that something is a cow) or . . . or (x is in physical state S′ and p = the proposition that *thinking* is a 2-place relation that holds between people and propositions). If, to avoid this circle, one omits the offending clause, the resulting definition would then fail to provide a necessary condition. (Note that, for an analogous reason, valuative properties cannot have infinitary disjunctive definitions in terms of their "naturalistic realizations." See also note 46.)

When a person thinks that he is thinking something, one and the same relation is involved twice—once in the content and once as the relation holding between the person and the content. This dual role is what undergirds the argument against Ramsified functionalism in section 1, for it licensed us to quantify (with one and the same quantified predicate variable) both the embedded and unembedded occurrences of psychological predicates. Initially, language-of-thought functionalism seemed to provide a systematic vehicle for avoiding these damaging quantifications. But these functionalists were then obliged to identify the content of Mentalese psychological predicates. They could succeed (without running into the circularity and levels problems) if those contents were grounded—somewhere in the hierarchies thinking_0, thinking_1, . . . ; desiring_0, desiring_1, . . . ; etc. But the standard mental relations are not grounded: on a given occasion, one and the same relation can be involved twice (once in the content and once as the relation holding between the person and the content), and whether a given mental relation plays this dual role on a given occasion is not determined by what nonpsychological propositions happen to be in the range of the relation. This phenomenon of ungroundedness is one of the very hallmarks of thinking, and the grounded-hierarchy conceptual role theory cannot capture it.

This leaves the conceptual role theory right where it was before the excursion into these hierarchies, namely, with a viciously circular analysis. (For example, if it is nomologically necessary that 'T' behaves in sentences in the respective Boxes (Thinking Box, Belief Box, etc.) in the way the predicate 'thinks' behaves in psychological theory $\mathcal{A}$, the content of 'T' $=_{\text{def}}$ the relation of *thinking*.) Is there any way to avoid the circle (besides retreating to Ramsification)? Evidently not. At least, every straightforward attempt fails. I will give one revealing illustration, namely, a recursive approach.

As with the hierarchy approach, suppose that the content-of relation has been defined for all nonpsychological Mentalese primitives; that would be the initial clause of an inductive definition. There would then be two sorts of inductive clause. One would concern the definition of the content-of relation for the various categories of complex expressions (existential generalizations, negations, etc.). The other would concern the definition of the content-of relation for Mentalese psychological primitives ('T', 'D',

etc.): the content of 'T' $=_{def}$ the relation holding between x and p such that a Mentalese sentence whose content is p is tokened in x's Thinking Box. And so forth. The previous remarks, however, show what is wrong with this approach. Ungrounded Mentalese sentences such as '(∃p) i T p' can be tokened in one's Thinking Box. Accordingly, the induction clause would fix the content of 'T' only if the content of '(∃p) i T p' were fixed earlier in the induction. But the content of '(∃p) i T p' would be fixed only if the content of 'T' were fixed still earlier in the induction. The inductive clauses thus fail to fix any of these contents: the vicious circle is not broken.

2.3 Retreat to Ramsification

The only move our functionalists have at this point is to retreat to some kind of Ramsification. But there is a dilemma here: Ramsification would indeed solve the circularity problem, but only at the price of reintroducing the levels problem. The simplest such Ramsification would be along the following lines: the content of 'T' $=_{def}$ the relation holding between x and p such that, for some first-order **R** satisfying A, x is related by R_2 to p. But, by the main argument of section 1, this definition would imply, contrary to ontological functionalism, that thinking must be a first-order relation. Our functionalists might propose to complicate this definition by explicitly incorporating the doctrine of language-of-thought functionalism into the psychological theory $\mathcal{A}$ and by including clauses that expressly require the indicated first-order **R** to be nomologically correlated with (relevant Mentalese sentences containing associated) Mentalese psychological predicates—respectively, 'P', 'T', ... But this would be a gratuitous complication, for as with the simpler definition, this definition would still imply that thinking is first-order.

As a last resort, our functionalists might revert to the kind of Ramsification, discussed at the close of section 1, that drops the explicit restriction on **R** to *first-order*. For example, with this explicit restriction dropped, the first Ramsified definition in the preceding paragraph would become: the content of 'T' $=_{def}$ the relation holding between x and p such that, for some **R** satisfying A, x is related by R_2 to p. Notice, however, that this definition would be correct

only if thinking could be defined directly by means of the sort of unrestricted Ramsified definition discussed at the close of section 1: x thinks p iff_{def} for some **R** satisfying A, x is related by R_2 to p. The argument at the close of section 1 then shows that the latter definition would be correct only if the standard mental properties were the *only* properties satisfying A. (As before, I am using 'property' for both properties and relations.) Therefore, the proposed definition of the content of 'T' would be right only if the values of its predicate variables were the standard mental properties themselves. This means that the proposed definition 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." Insofar as the proposed definition must quantify over the standard mental relations themselves, it endows them with an ontological primacy inconsistent with the basic functionalist picture. Moreover, as I just indicated, this definition of the content of 'T' would be right only if thinking has a direct Ramsified definition that makes no mention of language of thought. But we saw at the close of section 1 that this direct Ramsified definition has only ideological significance; from an ontological point of view, the definition is correct only if the standard mental properties are first-order. Thus, as far as this ontological issue is concerned, the excursion into language-of thought theory is a gratuitous complication.

These points hold *mutatis mutandis* for the more complicated kind of Ramsified definition mentioned two paragraphs above (that is, a Ramsified definition containing a clause expressly correlating **R** with tokenings of Mentalese psychological predicates) but now with the explicit restrictions on order removed. First, in this definition **R** must, as before, be the standard mental properties themselves. So the primary tenet of ontological functionalism is violated: the standard mental properties are not definable wholly in terms of the general pattern of interaction of ontologically prior "realizations"; by virtue of quantifying over the standard mental relations, the envisaged Ramsified definitions accord them an ontological primacy that functionalists reject. Second, once this point is granted, there is no reason to formulate one's Ramsified definitions in terms of language of thought; as before, language of thought makes no contribution. The simplest argument for this

appeals to results from section 3. There I will show that there are always families of deviant properties **R** that satisfy A and, hence, that the simple Ramsified definitions are mistaken. A wholly analogous argument can be given to show that the envisaged language-of-thought Ramsified definitions are likewise mistaken. The only promising way to save definitions of either kind is to restrict **R** to "natural" universals. Once this kind of restriction is imposed, however, it is completely implausible that the resulting restricted simple definition would be subject to counterexamples that are not also counterexamples to the resulting restricted language-of-thought definition.[34] Since it does nothing to block potential counterexamples, the excursion into language of thought is, as before, gratuitous.

3. Ideological Functionalism

Our conclusion is that ontological functionalism in each of its leading versions—Ramsified and language-of-thought—is mistaken. Contrary to ontological functionalism, the standard mental properties are first-order nonphysical properties. But this leaves it open whether "ideological" functionalism is feasible: although Ramsified definitions lack the ontological significance claimed by most functionalists, perhaps they correctly identify, in exclusively nonmental terms, *which* first-order nonphysical properties are the standard mental properties. The above discussion of the ungrounded nature of thinking, however, allows us to refute even this form of ideological functionalism.[35] The

[34]This is not to say that the definitions are correct: nothing I have said rules out the possibility that they are both subject to some further kind of problem, such as inverted spectrum. That is an independent question.

[35]This conclusion could be reached quite directly if one assumed the strong auxiliary view that every property is either a simple or else a complex formed in a unique way from simples and that a property is definable only if it is a complex. On this view, at most one Ramsified definition of a given property could be correct. The problem is that there are always a number of logically equivalent formulations of any given psychological theory $\mathcal{A}$. Given that thinking is self-embeddable and ungrounded, we can show that for any two such formulations of $\mathcal{A}$, the Ramsified definitions of thinking based on them would not be even materially equivalent. (To see why, see note 43.) Because there can be nothing in principle that would make one of these Ramsified definitions stand out as the "right one," the only reasonable conclusion to draw is that none of them would be correct. In the text I do not use this argument, for I want an argument that is

idea is to describe a system of nonstandard properties that behave with respect to each other and the external environment in exactly the same way that the standard mental properties behave with respect to each other and the external environment. Note that this style of argument, if correct, also refutes ontological functionalism (as well as the Armstrong-Lewis picture). It does not, however, show that the standard mental properties are ontologically first-order; the arguments of sections 1 and 2 are needed to show that. In what follows I will use 'functionalism' to apply to both ideological and ontological functionalism, and will use 'ideological functionalism' and 'ontological functionalism' only when it is necessary to distinguish the two.

To fill out the intuitive idea of the argument, let us engage in a bit of fancy. On Aristotle's conception, the unmoved mover necessarily contemplates its own contemplation and nothing else. For simplicity, let us represent this as follows, where u is the unmoved mover: Necessarily, u thinks that u thinks something, and u thinks nothing else. Consider this modal sentence to be a miniature psychological theory $\mathcal{A}$. Pretend that thinking is the only standard psychological attitude, and pretend that there could be no other thinking being besides u. Would the (unrestricted) Ramsified definition of thinking based on $\mathcal{A}$ be successful? No, not if there is a deviant relation thunking ($\neq$ thinking) such that: Necessarily, u thunks that u thunks something, and u thunks nothing else. Why? Because there would not be a *unique* relation satisfying the matrix A associated with $\mathcal{A}$; but, according to the argument at the close of section 1, the Ramsified definition would be correct only if there were a unique relation satisfying A. Given our (silly) pretenses, thinking could not be defined in terms of the general way it behaves with respect to itself and the external environment; every candidate functional definition would be powerless to distinguish thinking from thunking. (This includes language-of-thought functionalism, for there would be no way to identify thinking, as opposed to thunking, as the content of the Mentalese psychological predicate 'T'.) Of course, to use this idea in a serious argument against functionalism, we would need to drop the silly pretenses.

---

consistent with the (plausible) thesis that definable properties need not be complex.

Specifically, we would need to produce a whole system of deviant relations, one for each of the standard psychological attitudes, and these deviant relations would need to behave with respect to one another and the external environment in the way the standard psychological attitudes really do. We will be able to do this by means of a "diagonal" argument.

Functionalists should not take this argument any less seriously because it is "diagonal." In the last hundred years a number of substantive philosophical doctrines have been defeated with diagonal arguments (for example, by Russell, Tarski, Gödel, Church, Turing). When formulated carefully and precisely, functionalism is a highly technical view; it should be no surprise that technical ideas are needed for assessing it. The "diagonalization" in our argument will be achieved by means of "self-involving" intensions. This sort of intension has been vigorously studied in recent years under the general rubric of common (or mutual) knowledge.[36] There are intuitively compelling examples of such intensions, and there is fairly wide consensus that they play an important role in the explanation of intelligent cooperative behavior. There also exist a variety of formally consistent theories for dealing with them.

3.1 Argument from Self-Involving Intensions

Consider an example resembling Kripke's Watergate example.[37] Suppose that Nixon is taking notes about a man whom he is watching over closed-circuit television and whom he takes to be Dean. The man appears to be watching closed-circuit television and writing notes about what he is seeing. Nixon writes in his own notebook, "The sentence that man is writing would be worth reading." This sentence contains a definite description—'the sentence that man is writing'—that refers to the sentence the man is writing. Suppose that, as a matter of fact, the man Nixon is watching is not Dean but Nixon himself. No matter. The definite description 'the

[36]See, for example, Jon Barwise, "Three Views of Common Knowledge," in *Theoretical Aspects of Reasoning about Knowledge*, vol. 2, ed. M. Vardi (Los Altos, Calif.: Morgan Kaufmann, 1988), 365–80.

[37]Kripke, "Outline." I will consider an artificial one-person example; obviously there are analogous real-life examples involving chains of people.

sentence that man is writing' then refers to the sentence Nixon is writing; the sentence is about itself.

Instead of using the indicated definite description, Nixon might in the course of his notetaking use a demonstrative—say, 'that sentence'—to refer to the sentence the man on the closed-circuit television is in the process of writing. Alternatively, he might introduce a name—say, 'Plummer'—with the intention of using it to refer to the sentence that the man on the closed-circuit television is in the process of writing. For example, he might do this right in the course of his notetaking by writing, "Plummer would be worth reading." As long as he had the relevant intentions, 'Plummer' would refer to the sentence that the man on the television is writing. Since the man on the television is Nixon himself, 'Plummer' would refer to the very sentence he is writing. So, as in the earlier case, this sentence would be about itself.

Let us now shift the example from sentences to propositions. Suppose Nixon writes, "The proposition that man is asserting is unknown to Hoover." This sentence contains a definite description—'the proposition that man is asserting'—that refers to the proposition the man on the television is asserting. Since the man is Nixon himself, the proposition to which the definite description refers is the proposition Nixon himself is asserting. The proposition is about itself. Now, as before, a demonstrative or a name could be used instead of a definite description. For example, on a similar occasion Nixon could introduce a name—say, 'Nightshift'—to refer to the very proposition the man on the television is asserting. He might do this by asserting, "Nightshift is unknown to Hoover." As long as he had the relevant intentions, 'Nightshift' would refer to the proposition the man on the television is asserting. Since that man is Nixon, 'Nightshift' would refer to the proposition Nixon is asserting, namely, the proposition that Nightshift is unknown to Hoover. So we again have a proposition that is about itself. Of course, this proposition would be slightly different from the preceding one, for a sentence containing a definite description is not strictly synonymous with one containing a name in its place.

Just as there can be self-involving propositions, there can be self-involving attributes (and relations). Consider a variation on the above example. As before, Nixon is watching a man whom

he takes to be Dean but who is in fact Nixon himself. Taking the man to be writing notes about Mitchell, Nixon writes,

> Mitchell has the attribute of having each attribute which that man has attributed to him today, including the very attribute the man is attributing to him right now.

Given that the man is Nixon himself, the definite description 'the very attribute the man is attributing to him right now' refers to the attribute of having each attribute which that man has attributed to him today, including the very attribute the man is attributing to him right now.

Rather than using the description 'the attribute the man is attributing to him right now', Nixon could introduce a name—say, 'Password'—with similar effect. As long as he has the appropriate intentions, he might introduce 'Password' right in the course of writing,

> Mitchell has the attribute of having each attribute which that man has attributed to Mitchell today, including Password.

'Password' refers to the attribute which that man is attributing to Mitchell right then. But that attribute is the same one he refers to with the intensional abstract 'the attribute of having each attribute which that man has attributed to Mitchell today, including Password'. Thus, 'Password' refers to the attribute of having each attribute which that man has attributed to Mitchell today, including Password. Of course, this attribute would be slightly different from the attribute referred to with the intensional abstract in the preceding example, for the reference of an intensional abstract is not preserved when a constituent definite description is replaced with a name.

With these preliminaries in place, we are ready for an example bearing on functionalism. To simplify things, let 'Guilty' be a name of the proposition that Mitchell is guilty. Suppose Nixon believes that the man on the television whom he takes to be Dean is in the process of writing a sentence of the form 'Mitchell and Guilty stand in relation R'. Suppose that in a fit of philosophical verbosity Nixon writes,

> Mitchell and Guilty stand in the relation holding between x and p such that (for some q) x thinks q and p is just like q except that thinking and the relation which that man is in the process of ascribing are everywhere interchanged.

On analogy with our previous examples, the definite description 'the relation which that man is in the process of ascribing' would refer to

> the relation holding between x and p such that (for some q) x thinks q and p is just like q except that thinking and the relation which that man is in the process of ascribing are everywhere interchanged.

Of course, instead of using this definite description, Nixon might with similar effect introduce a primitive name, 'thunking', in the very act of his notetaking. As long as he has the appropriate intentions, he might do this in the course of writing,

> Mitchell and Guilty stand in the relation holding between x and p such that (for some q) x thinks q and p is just like q except that thinking and thunking are everywhere interchanged.

Given that Nixon had the relevant intentions, 'thunking' would refer to

> the relation holding between x and p such that (for some q) x thinks q and p is just like q except that thinking and thunking are everywhere interchanged.[38]

Let the predicate 'thunk' express this relation. To see what this relation is like, consider some examples. The proposition that 1 + 1 = 2 does not involve the affected relations (thinking, thunking). Therefore, x thunks that 1 + 1 = 2 iff x thinks that 1 + 1 = 2. For the same reason, x thunks Guilty iff x thinks Guilty. However,

---

[38]In symbols: the relation

$$\left[\text{xp: } (\exists q)\left(\text{x thinks q } \& \text{ p} = \text{q}\left[\frac{\textit{thinking, thunking}}{\textit{thunking, thinking}}\right]\right)\right].$$

the proposition that x thinks something does involve the relation of thinking. The result of interchanging thinking and thunking is the proposition that x thunks something. Consequently, x thinks that x thinks something iff x thunks that x thunks something. And so on for more complex cases.

Returning now to functionalism, suppose that each nonlogical constant in the psychological theory $\mathcal{A}$ is physical, mathematical, or psychological; and suppose that $\mathcal{A}$ is stated within some standard formulation of elementary predicate logic supplemented with modal operators and proposition abstracts (that is, 'that'-clauses).[39] Let us pretend, moreover, that 'thinks' is the only psychological predicate in $\mathcal{A}$. And let it be granted that thinking is not thunking. (I will return to this premise in a moment.) Finally, let $\mathcal{A}^*$ be just like $\mathcal{A}$ except that 'thunks' is everywhere interchanged for 'thinks'. Then, by a straightforward inductive argument, we can show that $\mathcal{A}$ and $\mathcal{A}^*$ are necessarily equivalent. It follows that thinking would satisfy A iff thunking also satisfies A (where A is the matrix formed from $\mathcal{A}$). Therefore, there is not a *unique* relation (that is, thinking) satisfying A. By an argument like that at the close of section 1, however, the Ramsified definition based on $\mathcal{A}$ would be correct only if there were a *unique* relation satisfying A. So the Ramsified definition based on $\mathcal{A}$ is mistaken, and the associated version of functionalism fails.[40]

The foregoing remarks can easily be generalized from a single relation (thinking) to the full list of standard mental relations (thinking, desiring, . . .), enabling us to reach the same conclusion for them.[41] Moreover, by analogous arguments we can reach the

[39]Most functionalists should be happy with this logical framework: its quantified variables may range freely over properties, relations, and propositions, as well as particulars, and it is fully equipped to give standard propositional-attitude reports.

[40]As indicated earlier, language-of-thought functionalism would be undermined for a similar reason: there would be no way to identify thinking (as opposed to thunking) as the content of the Mentalese psychological predicate 'T'.

[41]In this argument, we would have a sequence of (simultaneously introduced) nonstandard relations thonking, dosiring, ... The nonstandard relation thonking = the relation

$$\left[\text{xp: } (\exists \text{q})\left(\text{x thinks q } \& \text{ p} = \text{q}\left[\frac{\textit{thinking, desiring, . . ., thonking, dosiring, . . .}}{\textit{thonking, dosiring, . . ., thinking, desiring, . . .}}\right]\right)\right].$$

Likewise for the nonstandard relation dosiring. And so on. Grant that

same outcome when $\mathcal{A}$ has a richer background logical framework.[42] We are thus led to the general conclusion that functionalism based on the associated Ramsified definitions is mistaken.

3.2 Thinking Is Not Thunking

The above argument uses the premise that thinking $\neq$ thunking. I will now motivate this premise. (In the generalized version of the argument, the corresponding premise would be that thinking $\neq$ thonking; desiring $\neq$ dosiring; . . . . The reasons hold *mutatis mutandis* for this more complex premise.)

The most serious argument is theoretical. But first some intuitive points: On its face, thunking just *seems* different from thinking. My intuition about this is quite vivid. Now some people tell me that they lack intuitions in this area. But surely it seems to you that thunking is a "Cambridge" (concocted) relation, whereas it does not seem to you that thinking is such a relation. Perhaps those who, in spite of this, continue to doubt that thinking is different from thunking fail to recognize how slight the difference need be to be significant. If they are different in any way, however slight it might seem, they would be very different; indeed, they would not even be materially equivalent.[43] For another intuitive consideration,

---

thinking $\neq$ thonking; desiring $\neq$ dosiring; . . . . Then, if $\mathcal{A}$ is otherwise as above, we can show: the sequence of standard mental relations would satisfy A iff this sequence of nonstandard relations also satisfies A. So the Ramsified definitions based on $\mathcal{A}$ would be mistaken, and the associated version of functionalism would fail. Note that if $\mathcal{A}$ deals with the relation of referring, the latter should be deemed a psychological relation and a nonstandard relation of roferring would also be included in the list of nonstandard relations.

[42]For example, in a similar but more complex fashion, one could conversationally introduce the name of an operation * with the following features. *(thinking) $\neq$ thinking; *(desiring) $\neq$ desiring; etc. For an arbitrary complex intension w (either a proposition, a complex property, or a complex relation), *(w) is the complex intension that arises from w by replacing each simple constituent u of w with *(u). And for simple properties u, *(u) = the property of being a y such that, for some z, z is an instance of u and y = *(z). In symbols: *(u) = [y: $(\exists z)$(z is an instance of u & y = *(z)]. Likewise for simple relations. Then, on the assumption that thinking, desiring, etc. and their *-counterparts are simple relations, we can show that p and *(p) are necessarily equivalent for arbitrary propositions p. (Alternatively, if they are complex, we can reason as we did in note 35.)

[43]This follows from the way the two relations would show up in embedded propositions. To illustrate: Suppose that thinking and thunking differ

here is a further example. Suppose that over closed-circuit television Yeltsin is secretly watching someone whom he takes to be Gorbachev, but who is Yeltsin himself. He takes the person to be writing something about Shevardnadze. In the course of his notetaking, Yeltsin introduces the name 'plotnost' in such a way that it refers to the following: the relation holding between x and p such that (for some q) x thinks q and p is just like q except that thinking and *plotnost* are everywhere interchanged. I find it quite unintuitive that plotnost and thunking would be identical. Once again, the difference need only be slight. But if plotnost and thunking are not identical, at least one of them must be distinct from thinking. Given that at least one of them is distinct from thinking, however, it would be utterly mysterious if both were not. For in all relevant respects each is exactly as unlike thinking as the other. Thus, we are led to the conclusion that thunking is distinct from thinking.

Now for the theoretical argument, which most people find very compelling: Recall the sort of self-involving propositions with which we began this section. Liar-paradox propositions belong to this family. In our original Nixon example, Nixon might have written, "The proposition that man is asserting is false." The definite description 'the proposition that man is asserting' would refer to the proposition Nixon would be asserting; so that proposition would be about itself and, as such, would be subject to the usual reasoning that leads to paradox. Alternatively, he could have introduced a name—say, 'Scapegoat'—for the proposition he took the man on the closed-circuit television to be asserting. He could do this by writing, "Scapegoat is false." If he had the relevant intentions, 'Scapegoat' would refer to the proposition that Scapegoat is false. Now comparable situations could arise any number of times for any number of people. Kripke might introduce the name 'John' to refer to a proposition he has in mind, namely, that John

---

in some way, however slight it might seem. If so, it would be possible to think (consciously and explicitly) a proposition involving thinking and not at that time to be thinking (consciously and explicitly) the corresponding proposition involving thunking. For example, I am now thinking that I am thinking, but I am not now thinking that I am thunking. But I think that I am thunking iff I thunk that I am thinking. Since I am not now thinking that I am thunking, it follows that I am not now thunking that I am thinking. Thus, the proposition that I am thinking is not in the range of the thunking relation. But it is in the range of the thinking relation. So the two relations are not materially equivalent.

is false. Tarski could introduce the name 'Jerzy' to refer to a proposition he has just come upon, namely, that Jerzy is false. And on and on. Now upon considering the question, it *just seems* to me that Scapegoat and John would be different somehow and, likewise, that John and Jerzy would be different somehow. This is my immediate, untutored response. Nearly everyone has this response when they put aside their theoretical commitments and general inclination to avoid taking stands. Those of us who have this intuitive response are committed to accepting it at face value unless there are independent, non-question-begging reasons to do otherwise. But there are no such reasons. On the contrary, if Scapegoat, John, Jerzy, etc. were all identical, this would amount to holding that there can be only *one* liar proposition of the indicated form! Evidently, no one has the intuition that this is so. And, as far as I know, no one working on the truth paradoxes holds—or has any reason to hold—that there must be exactly one liar-paradox proposition of this form.[44] (These logicians certainly would not let their solution depend on the dubious assumption that there is only one such proposition.) Thus, we are led to accept our intuition at face value.

Let us agree, then, that there is a plurality of self-involving liar-paradox propositions of the indicated form. In this case, the same thing should hold for paradoxical *attributes*. For example, with appropriate intentions, Nixon could introduce a name 'Conundrum' to refer to the attribute of not having Conundrum as an attribute. Likewise, Russell could introduce 'Enigma' to refer to the attribute of not having Enigma as an attribute. And so forth. Given that the self-involving propositions Scapegoat, John, Jerzy, and so forth are distinct, uniformity supports the thesis that these self-involving attributes Conundrum, Enigma, and so forth are also distinct. Now if these self-involving paradoxical attributes are distinct, uniformity also supports a further generalization (for any fixed ⌜ . . . ⌝ of the relevant sort): there could be any number of appropriately intro-

---

[44]In *Non-well-founded Sets*, CSLI Lecture Notes, no. 14 (Stanford: CSLI, 1988), Peter Aczel considers an axiom for set theory that implies that there is exactly one set x such that $x = \{x\}$. As far as I can see, there is no convincing support for this axiom. In any case, this axiom is not relevant to the point under discussion in the text, for the axiom is concerned with extensional entities, whereas we are concerned with intensional entities. Presumably, intensionality makes all the difference.

duced names α such that the intensional abstracts ⌜the attribute of being something such that . . . α . . . ⌝ would refer to distinct attributes. And given this, uniformity supports the analogous conclusion for self-involving *relations*. But the relations thunking and plotnost are just such relations. So they too must be distinct. But we saw a moment ago, if thunking and plotnost are distinct, at least one of them must be distinct from thinking. Given that at least one of them is distinct from thinking, it would be utterly mysterious if both were not. For in all relevant respects each is exactly as unlike thinking as the other. So, in particular, thunking is distinct from thinking.

Although the foregoing does not *prove* that thunking and thinking are distinct, it has significant persuasive power. Advocates of the indicated Ramsified definitions of mental properties turn out to have surprising commitments in *logic*, namely, to the thesis that there can be at most one liar proposition of the form discussed above! This thesis conflicts with our intuitions and has absolutely no independent support. Thus, advocates of such definitions are in an epistemically untenable situation. Hardly what one would want from the leading philosophy of mind.

3.3 Conclusion

In view of all these considerations, I conclude that our "diagonal" argument stands. Faced with this, ideological functionalists have only one option (besides simply abandoning their view), namely, to revise their Ramsified definitions in a certain ontologically significant way. As we have noted, thunking, plotnost, etc. are intuitively "Cambridge" entities, not genuine "natural" universals. Suppose that the standard mental properties are genuine "natural" universals (as indeed they intuitively seem to be). The proposal would be to revise the ideological functionalists' Ramsified definitions by explicitly requiring **R** to be genuine "natural" universals. The resulting definitions would then be immune to the above sort of counterexample.[45]

[45]Ontological functionalists may also make this move, holding that the standard mental properties are "natural" second-order universals. Although the resulting view would then avoid the argument of this section, it would still be mistaken, for sections 1 and 2 show that the standard mental properties are first-order, not second-order.

There is thus only one way in which ideological functionalism might be true: the standard mental properties—which have already been shown to be first-order and nonphysical—would have to be "natural" universals as well. But the same conclusion would hold if ideological functionalism were false. For if ideological functionalism were false, there would be no alternative but to take the standard mental properties to be in principle undefinable. At the same time, they play a part in our best overall theory, and properties that play a part in our best overall theory but are in principle undefinable must be considered genuine "natural" universals. So either way, we are led to the conclusion that the standard mental properties are first-order nonphysical "natural" universals. Mental properties are *sui generis*. No vestige of reductionism survives.[46]

University of Colorado at Boulder

The reflections in the text suggest another fault with the Armstrong-Lewis picture. Suppose, as their doctrine requires, that there are first-order physiological properties that satisfy A. Then, by a "diagonal" argument somewhat similar to that in the text, we can show that associated with them there would have to exist any number of deviant first-order properties also satisfying A. It follows that the Armstrong-Lewis Ramsified definitions would be mistaken. To avoid this problem, Armstrong and Lewis might try to mimic the strategy in the text by restricting the first-order physiological "realizations" satisfying A to "natural" universals. But it is wholly implausible that there are "natural" first-order physiological universals of the envisaged sort. Think of what it would be for first-order "realizations" **R** to behave with respect to one another in a fully A-like way. R_2 (the "realization" of thinking) would need to hold between subjects and fine-grained propositions and would need to be ungrounded (for example, it should be possible that x R_2 the proposition that, for some p, x R_2 p, and it should be possible for this to occur independently of whether x R_2 any other proposition). And now we are being asked to believe that R_2 is a "natural" universal belonging to an actual physical science, namely, physiology! Certainly, physiologists would never have reason to posit the existence of a basic physiological relation meeting these remarkable conditions. Anyone who seriously posits such a basic physiological relation surely is only "spreading the mind" onto the brain in an unscientific manner. (Corresponding posits of new "natural" physical properties by computer enthusiasts would likewise be unscientific "spreading of the mind" onto physical machines.)

Notice, incidentally, that if independent intuitive considerations show that the standard mental properties are "natural" universals, then the last remarks would lead to a new style of refutation of the identity thesis itself.

[46]If correct, the considerations in this paper also seem to undermine efforts to reduce valuative properties to nonvaluative properties by means of Ramsification, for valuative properties exhibit a self-embeddability akin to that exhibited by mental properties.