6

The Self-Consciousness Argument: Functionalism and the Corruption of Content

George Bealer

What exactly is the relationship between physical and mental properties?* On this question materialist philosophers and cognitive scientists have a very limited range of basic alternatives (short of rejecting mental properties as real properties; cf. Churchland 1981). Broadly speaking, these alternatives are: some form of behaviorism, some form of the identity theory,[1] and some form of functionalism. In recent years, however, compelling criticisms have been made against the various forms of behaviorism and of the identity theory—leaving functionalism, in one or another of its many guises, as the most promising alternative upon which philosophers and cognitive scientists can pin their materialist aspirations.

Today, functionalism unifies and animates much of materialist philosophy of mind and cognitive science, at least as its tacit conceptual framework. Broadly functionalist sentiments have been articulated at one time or another by a long list of notable philosophers and cognitive scientists. Here is a small sampling of the philosophers: Lewis (1966), Putnam (1970), Harman (1973), Fodor (1981), Lycan (1987), Shoemaker (1981), Jackson (2003). And a small sampling of the cognitive scientists: Pylyshyn (1984), Gardner (1985), Minsky

* In writing this paper I benefitted from numerous insightful comments and suggestions from Iain Martel, John Bengson, and especially Marc Moffett. I also benefitted from a lengthy correspondence with Sydney Shoemaker on the Self-consciousness Argument. My thanks to Leslie Wolf for his meticulous work on the proofs.

[1] I will take the identity theory to be the doctrine that standard mental properties (thinking, being in pain, etc.) are identical to first-order physical properties (e.g., being in such and such neural state); I give my reasons for rejecting the identity theory in Bealer (1994; see also note 26 below). Here and in what follows I mean by standard mental properties the sort designated by canonical gerundive phrases 'thinking', 'being in pain,' and so forth (for more on this point, see note 3). As a terminological convenience, here and elsewhere I use 'property' for both properties and relations; I will use 'relation' when the context requires. It turns out that mental relations will be central to the debate.

(1985), Chomsky (1988), Johnson-Laird (1988), Newell (1990), Jackendoff (1992; 1997). Because of functionalism, a great many philosophers and cognitive scientists now acquiesce in the belief that there are no serious conceptual obstacles to a broadly materialist understanding of the mind and that the remainder of the story is largely empirical. To put the point another way, since the cognitive revolution (since our renewed efforts to peer into the "black box"), cognitive scientists are once again willing to implement robust realist mental notions in their theories—but to no small degree because functionalism has been thought capable, at least in principle, of explaining those notions within a broadly materialist framework. In this way, functionalism has served to clear the intellectual conscience of philosophers and cognitive scientists seeking a materialist understanding of the mind.

If, therefore, there is a principled barrier to functionalism, this vision of materialist philosophy of mind and cognitive science is put in jeopardy. In this chapter I argue that there is such a barrier created by self-conscious intentional states—conscious intentional states that are about one's own conscious intentional states. As we will see, however, this result is entirely compatible with a scientific theory of mind, and, in fact, there is an elegant non-reductive framework in which just such a theory may be pursued.

1. WHAT IS FUNCTIONALISM?

Functionalism has earned its role by promising to resolve, in the words of Jerry Fodor, a "nasty dilemma" in the materialist program in cognitive science—specifically, by preserving the virtues of behaviorism and the mind–brain identity theory while disavowing their shortcomings. In his well-known *Scientific American* article Fodor (1981) summarizes the historical situation thus:

> On the one hand the identity theorist (and not the logical behaviorist) had got right the causal character of the interactions of mind and body. On the other the logical behaviorist (and not the identity theorist) had got right the relational [dispositional] character of mental properties. Functionalism has apparently been able to resolve the dilemma.

(It should be emphasized that Fodor's "computational language-of-thought" functionalism is only one of a wide spectrum of functionalist theories. Nothing in this chapter turns on accepting Fodor's version.)

The identity theory is thought to be right in that it treats mental properties (states) as real properties having genuine causal efficacy. But it is thought mistaken because it identifies mental properties with *particular* physiological properties, whereas plainly nothing prevents them from being realized in a multiplicity of different ways from species to species and perhaps even from individual to individual. For example, the property of being in pain can be realized by firing "C-fibers" in one sort of creature, firing "D-fibers" in another,

etc. (where 'C-fiber,' etc. are being used as mere *dummy terms* for whatever properties physiology ultimately settles upon). Intuitions of this sort are simply overwhelming in the case of intentional properties, for example, the intuition that intelligence, knowledge, and thought can be realized in a multiplicity of different ways in different species. It would be completely unscientific and ad hoc to insist that this is not so. The functionalist response is that what is common across species are, not first-order physiological properties, but rather the functional roles of such properties.

Behaviorism, on the other hand, is thought to be right in that it treats mental properties (states) as having an essential dispositional character, where those dispositions are ultimately anchored in observable inputs and outputs. But it is thought to be incorrect in that it treats mental states as being fixed *entirely* in terms of inputs and outputs, disregarding their essential interaction with one another. That interaction, however, is thought to be required to account for the fact that two organisms could in principle have the same input–output functions and yet have different mental states. And treating mental states as physically realized internal states is thought to be required to explain rational behavior (resulting from means–ends reasoning) and the sort of peculiarly human linguistic phenomena celebrated by Chomsky (1959) and many others.

Against this historical background, Fodor (1981) states:

> [The cognitive sciences] have in common a certain level of abstraction and a concern with systems that process information. Functionalism, which seeks to provide a philosophical account of this level of abstraction, recognizes the possibility that systems as diverse as human beings, calculating machines, and disembodied spirits could all have mental states. In the functionalist view the psychology of a system depends not on the stuff it is made of (living cells, mental or spiritual energy) but on how the stuff is put together.

An account of what mental properties are falls naturally out of this view.

According to functionalism, thinking, being in pain, and other such mental properties are *second-order*: they consist in there being *other* properties, namely, first-order *realizations* that have appropriate interactions with one another and the external environment.[2] Accordingly, functionalists hold that such mental properties can be defined wholly in terms of this general pattern of interaction of their realizations (the PATTERN, as I shall call it for brevity).[3] Because there

[2] For a gloss on 'first-order' and 'second-order,' see Putnam (1970). For now, suffice it to say that first-order properties are properties definable in terms of *specific* primitive properties of individuals. Second-order properties are not first-order properties but are definable by quantifying over (i.e., by speaking *generally* about) first-order properties. For example, the property of being square is a first-order geometric property, and the property of having *some* first-order geometric property or other is a second-order geometric property. These notions of first- and second-order underlie the ramified type theory presented in *Principia Mathematica*.

[3] This characterization of functionalism clearly fits standard "American" functionalism (advocated at one time or another by Putnam, Fodor, Block, Harman, Shoemaker, Loar, Lycan, Cummins, and many others). It also fits "Australian" functionalism (advocated by Armstrong, Lewis, Jackson,

can be more than one system of realizations that fits the PATTERN, functionalism is compatible with the intuition that a given mental property—say, being in pain or thinking—can be realized in a multiplicity of different ways. What such realizations have in common is precisely that they fit the PATTERN in the way definitive of being in pain, thinking, and so forth. (See, e.g., Block (1990) for further elaboration.)

The legacy of behaviorism is that behavioral input–output relationships provide the observable anchors in the PATTERN. The legacy of the identity theory is that the first-order physiological realizations underwrite the flow of causes and effects within the PATTERN. At the same time, this entire picture is, by design, consistent with the powerful collateral thesis that the internal transitions within the PATTERN are *computational* in character. Although a number of *philosophical* functionalists remain neutral on this computational requirement, a great many *scientific* functionalists embrace it and, indeed, use it to guide their (traditional or connectionist) research. What I have to say will apply equally to both versions of functionalism—explicitly computational and computationally neutral.

No doubt the major conceptual attraction of functionalism is that, if correct, it provides a very ingenious solution to the Mind–Body Problem. How? By making the relation between physical and mental properties completely transparent and unmysterious. For, when functional definitions are substituted into the following:

If a thing has some system or other of first-order physical realizations that fit the PATTERN, then it has the associated mental properties as well.

Braddon-Mitchell, Pettit, and others). True, the latter functionalists hold that *pain* is contingently identical to a certain first-order realization, namely, the occupant of the "pain-role" (say, firing C-fibers). But Lewis (1966) says, "I take 'the attribute of having pain' . . . as a *non-contingent* name of that state or attribute Z that belongs, in any world, to whatever things have pain in that world…." By parity, he, and his fellow Australian functionalists, would need to hold that the property of *being in pain* is necessarily identical to the property of having a first-order property which is an occupant of the pain-role. After all, like the gerundive phrase 'having pain,' the gerundive phrases 'being F,' 'having G,' etc., are also *non-contingent* names (i.e., they are rigid designators: they designate the same thing in counterfactual situations as they do in actual situations). Therefore, being in pain would have to be a second-order property definable in terms of the PATTERN of first-order realizations, just as in the general formulation in the text. (Note that, since being in pain is multiply realizable, it would not be identical to any one first-order physical property (e.g., having firing C-fibers). Therefore, the identity theory—as it is glossed in note 1—would be mistaken on Australian functionalism. I will return to this fact in the course of the argument.)

In this chapter, I will be primarily concerned with propositional-attitude psychology and intentional relations. These relations are expressed by standard propositional attitude verbs (e.g., 'thinks') and are designated by the associated canonical gerundive phrases ('the relation of thinking' or 'the thinking relation' or simply 'thinking'). The Self-consciousness Argument will focus on conscious intentional relations (and, in particular, the thinking relation). The argument will employ the premise that functionalism is committed to the following tenets: (1) the indicated gerundive phrases ('thinking,' etc.) rigidly designate second-order relations (thinking, etc.); and (2) there are associated first-order physical relations, which together with first-order physical properties, fit the PATTERN. Lewis commits himself to these two tenets, and so, it seems, do Armstrong, Jackson, and the other Australian functionalists.

the result is a tautology:

If a thing has some system or other of first-order physical realizations that fit the PATTERN, then there is some system or other of first-order physical realizations that fit the PATTERN and the thing has them.

In this sense, the body–mind relationship may be viewed entirely as a matter of logic and definitions (i.e., as analytic, in Frege's sense).[4] No puzzling non-logical (synthetic) relations need to be posited; at the same time, behaviorism and the identity theory are avoided. Moreover, functionalism is compatible with the following elegant form of materialism: even though mental properties might in principle be realized non-physically (is anyone certain that they cannot be?), they are in fact realized only physically; and, therefore, (given the truth of functionalism) all properties that are instanced in actual individuals are either first-order physical properties or higher-order properties that are logical consequences of them.[5]

Functionalism has many virtues, but there are also many prominent objections: anomalism (Davidson 1968); subjectivity and "what-it's-like" (Nagel 1974; Jackson 1986); inverted-spectra, absent qualia, Chinese nation, homunculus head (Block 1978); externalism and anti-individualism (Putnam 1975, Burge 1978); Searle's Chinese Room (Searle 1980); eliminativism (Churchland 1981); utopianism (Putnam 1988); consciousness and zombies (Chalmers 1996). I believe, however, that none of these is conclusive—that, at the very least, each can be rendered moot by a clever functionalist. Indeed, functionalism might well be wholly successful were it not for a central type of conscious intentional state, namely, *self-conscious thought* (thinking that one thinks *q*, desiring that one desires *q*, thinking that one desires *q*, being self-consciously aware that one feels pain, etc.). For this reason, I believe that self-conscious thought poses the most

[4] According to "psychofunctionalism" (Block 1978) these definitions must be discovered by empirical science and so are not a priori. And whether a given thing actually has a given system of first-order physical realizations is taken to be a contingent fact. Note that, both here and in what follows, 'physical realization' (and 'physical property') may be understood widely so as to include relevant physical facts about the external environment and perhaps even physical laws themselves.

[5] This is not to say that these higher order properties are physical *by nature*. For our purposes, a second-order property is physical by nature only if it can be expressed by some second-order formula in which all talk of properties is explicitly restricted to first-order *physical* properties. A great many functionalists (e.g., Putnam, Lewis, Fodor, etc.) reject the idea that mental properties are physical by nature, for they accept that it is possible for mental properties to be realized nonphysically. Accordingly, in their second-order definitions of mental properties, they would prohibit restricting the relevant quantifiers to first-order physical properties.

There is nevertheless a weak sense in which a higher-order property may be deemed physical. Suppose that in a typical second-order functional definition of a mental property, talk of properties is not explicitly restricted to first-order physical properties. And suppose that the first-order realizations of the mental property turn out to be physical as a contingent fact. Then, functionalists might deem the mental property to be *physical as a contingent fact*. Throughout this chapter, when I speak of physical properties, I will not mean physical in this weak sense but rather *physical by nature*. This is the dominant use of 'physical' in the literature.

formidable threat (and the only rigorous internal threat) to functionalism on all its familiar formulations.

The threat posed by self-conscious thought is especially apt today in light of the flurry of interest in the experiential aspects of consciousness (for example, Block 1978, 1995; Lycan 1987, 1997; Baars 1988, 1996; Crick and Koch 1990; Strawson 1994; Chalmers 1996; Jackson 2003). There is a growing tendency to think that all problems of intentionality (believing, thinking, desiring, deciding, etc.) have already been solved by traditional functionalism and that the only "hard problem" results from conscious experience (Strawson 1994; Chalmers 1996).[6] In this connection, consciousness is often actually equated with phenomenal experience, neglecting what traditionally (as far back as Descartes) was wholly central to consciousness—namely, conscious thinking (as in the *Cogito*). Plainly, conscious thinking and conscious experiencing have something in common, namely, consciousness itself. And this is no pun. The idea (Block, Strawson, Chalmers) that conscious thinking has a reductive explanation and conscious experiencing does not (or vice versa) yields an implausibly fragmented picture of consciousness (and, indeed, of the mind and the self).

The problem of self-conscious thought may be put as a dilemma. Either the standard functional definitions admit the wrong sorts of things as typical contents of one's conscious thoughts about one's current conscious states (since those contents would have to be propositions involving first-order *realizations* rather than mental properties themselves), or else the definitions are circular and so do not even count as definitions. The only way out of this dilemma is to abandon the primary tenet of functionalism (that mental properties can be defined wholly in terms of the PATTERN of their realizations) and to replace the standard reductive functional definitions with non-reductive counterparts. But doing this, we shall see, undermines functionalism's explanation of the relation between physical and mental properties and in turn its solution to the Mind–Body Problem itself.

2. FUNCTIONAL DEFINITIONS AND SELF-CONSCIOUS THOUGHT

In what follows I try to be faithful to the formulations of functionalism in the published literature, and I believe I succeed at this. If the argument is correct, those formulations of functionalism contain a serious internal difficulty. Functionalists therefore need to find a way to *revise* their view. I believe that this cannot be done without violating the central tenets and aims of functionalism.

[6] Galen Strawson introduced the term 'The Hard Problem' in his 1994 book. David Chalmers uses the term in his 1995 paper "Facing Up to the Problem of Consciousness" and in his 1996 book.

The most clear and precise formulations of functionalism are those based on the idea of "Ramsification."[7] Here a whole theory is converted into a definition by replacing its 'theoretical' (in this case, psychological) predicates with variables bound by the existential quantifier 'there exist.' As noted, psychological theory describes the PATTERN of interaction of the standard mental properties and relations with one another and the external environment. Let A be a comprehensive psychological theory specifying the PATTERN. $\boldsymbol{A}$ results from A by replacing psychological predicates with associated predicate variables 'R_1,' 'R_2,' . . . . Let '$\boldsymbol{R}$' be short for '$R_1, R_2, \ldots$.' Then, assuming that 'is in pain' is the first psychological predicate occurring in A and 'thinks' the second, functionalists then propose the following standard functional definitions:

x is in pain iff$_{\text{def}}$ there exist first-order realizations $\boldsymbol{R}$ satisfying $\boldsymbol{A}$ and x has R_1.

x thinks q iff$_{\text{def}}$ there exist first-order realizations $\boldsymbol{R}$ satisfying $\boldsymbol{A}$ and x is related by R_2 to q.

Consider the second definition. We know that the thinking relation is characterized by a number of quite distinctive interactive principles. For example, the following *Self-intimation Principle*: if a person is thinking something and engaging in introspection, he or she will think that he or she is thinking something. Perhaps qualifiers need to be added—for example, 'engaging in thorough and attentive introspection,' '*ceteris paribus*,' 'probably.' (If you prefer, 'is thinking something' could be replaced with 'is in pain,' 'is sensing red,' etc., in the antecedent and in the embedded occurrence in the consequent. But it is convenient to stick with 'is thinking something' because self-embedded attitudes—and also cross-embedded attitudes such as thinking that you are desiring—will prove to be of special interest later on.) The point is that some such principles, with or without qualifiers, would belong to psychological theory A, given that A is comprehensive. As Sydney Shoemaker (1994: 59) says, "[I]n many cases it belongs to the very essence of a mental state (its functional nature) that, normally, its existence results, under certain circumstances, in there being such awareness of it." David Lewis (1966) expresses much the same point thus:

> [Functionalism] allows us to include other experiences among the typical causes and effects by which an experience is defined. It is crucial that we should be able to do so in order that we may do justice, in defining experiences by their causal roles, to the introspective accessibility which is such an important feature of any experience. For the introspective accessibility of an experience is its propensity reliably to cause other (future or simultaneous) experiences directed intentionally upon it, wherein we are aware of it.

[7] Ramsey (1931: 212–36) proposed a technique, not for defining "theoretical" terms, but rather for eliminating them by means of existentially quantified predicate variables. To my knowledge, 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 (1966: 1–13). Martin's idea or variants on it were subsequently advocated by Putnam (1970), Lewis (1970), and a long list of others (Harman, Loar, Shoemaker, Block, Cummins, Jackson, etc.).

For simplicity, suppose that A_0 is a conjunction of some complex clause Q and the above Self-intimation Principle. The formula $\boldsymbol{Q}$ results from Q by replacing psychological predicates with predicate variables as before. Assume that 'introspects' is the third psychological predicate occurring in A. Then, stated in greater detail, the above standard functional definition of the thinking relation would be:

x thinks q iff_{def} there exist first-order realizations $\boldsymbol{R}$ such that (i) they satisfy $\boldsymbol{Q}$; (ii) if x R_2s something and x R_3s, then x will be related by R_2 to the proposition that he R_2s something; and (iii) x R_2s the proposition q.

Clause (ii) results from the Self-intimation Principle by replacing 'thinks' with 'R_2' and 'introspects' with 'R_3.' (To see that this is intensionally correct, see below.)

We can now pinpoint the problem: this functional definition implies that first-order realizations of the thinking relation (rather than the thinking relation itself) would be among the typical contents of our everyday self-conscious thoughts. To see why, suppose x is both thinking something and engaging in introspection. Then, by the left-to-right directions of the functional definitions of thinking and introspecting, there would be first-order realizations $\boldsymbol{R}$ which satisfy $\boldsymbol{A}$ such that: x R_2s something and x R_3s. Since this conjunction is the antecedent of clause (ii) and since (given that $\boldsymbol{R}$ satisfies $\boldsymbol{A}$) R_2 and R_3 satisfy clause (ii), it follows by modus ponens that: x is related by R_2 to the proposition that he R_2s something. But the right-to-left direction of the definition of thinking implies that, if x is related by such a first-order realization R_2 to an arbitrary proposition q, then x thinks q. So, given that x is related by R_2 to the proposition that he R_2s something, it follows that x thinks that he R_2s something. But R_2 is not the relation of thinking (i.e., the relation expressed by the predicate 'thinks' and denoted by the associated gerund 'thinking'; see note 3), which according to functionalism is a second-order relation; rather, R_2 is a first-order physical realization of the thinking relation. The upshot is that the functional definition admits the wrong sorts of things into the contents of our everyday self-conscious thoughts.[8]

One response to this argument is to "bite the bullet," that is, to hold that propositions involving such first-order physical realizations really are typical objects of the thinking relation. But this is wholly implausible once it is realized that we are talking about the relation of conscious explicit thinking. This is a highly *focused* relation. When a person is consciously and explicitly thinking that he or she is thinking something, typically the person will not be consciously

[8] Another possibility is that there simply are *no* first-order realizations that display the sort of self-embeddability characteristic of mental relations. In this case, functionalists would be committed to holding that there is *no* sequence of first-order realizations $\boldsymbol{R}$ satisfying $\boldsymbol{A}$. If so, the right-hand side of the definition would be null and therefore would not correctly define the thinking relation. Hence, it only helps functionalists to suppose that there are first-order realizations $\boldsymbol{R}$ satisfying $\boldsymbol{A}$.

and explicitly thinking *two* propositions, one involving the relation of thinking (i.e., the relation expressed by the predicate 'thinks' and rigidly denoted by the associated gerund, cf. note 3) and the other some first-order physical realization of the thinking relation, say, R_2.

Would it make sense to reply that, although there is indeed just one of these propositions you are thinking when you are thinking that you are thinking something, it is really the proposition involving a first-order realization, rather than the thinking relation itself? More specifically, is it really the proposition that you R_2 something? Obviously not. Indeed, the former proposition does not even entail the latter: after all, it is possible that someone think something even if no one bears R_2 to anything. Look at the question this way. Suppose a creature in a species with a different physical make-up is thinking that it thinks something. Then, by a simple one-step existential-generalization inference, both you and the creature could arrive at a single proposition on which you *agree*, namely, that *someone* thinks something. But this obvious possibility would be out of reach if the respective propositions you and the creature were originally thinking involved, not the thinking relation itself, but instead *distinct* physical realizations of the thinking relation (your R_2 and the creature's S_2).[9]

The most common worry about our main argument concerns intensionality—specifically, the fact that the embedded occurrence of 'thinks' in the Self-intimation Principle was replaced with an existentially quantified predicate variable. Was it right to do that?[10] Yes, but before explaining why, let me put the argument in the form of a dilemma for functionalists: either they accept that the pivotal existential generalizations are valid or they do not. If they do, then some version of the argument goes through. If they do not, then the resulting functionalist "definition" of thinking will not even qualify as a definition, for the undefined psychological expression 'thinks' would still occur on the right-hand side.[11] Thus, functionalists may go one of two ways on the intensionality issue, but whichever way they go, their Ramsified definitions are unsatisfactory.

[9] In view of these considerations, it should be clear that our main argument applies against both American and Australian functionalism. Each implies that, when you are thinking that someone thinks something, the proposition that you are thinking involves a first-order realization, rather than the relation of thinking itself. Likewise for the creature in the other species. Consequently, on both versions of functionalism, you and the creature would not agree that someone is thinking something. Some Australian functionalists might reply that the creature—suppose it is a Martian—is not thinking (but rather is M-thinking). It is a truism, however, that intelligence requires thinking well. Hence, if the Australian functionalist were correct, it would follow that the creature is not intelligent. But this is absurd: the existence of extraterrestrial intelligence cannot be disproved so easily. It follows, therefore, that Australian functionalism is mistaken.

[10] For more extended discussion of the intensionality issue, see Bealer (1997 and forthcoming b). Note that Lewis would agree that there is no intensionality problem (see Lewis 1972: note 8).

[11] Definitions of the following sort escape both horns of this dilemma:

The relation of thinking $=_{\text{def}}$ the unique relation T such that, necessarily, (for all x and q) $x\ T$s q iff for some first-order properties $\boldsymbol{R}$, (i) $\boldsymbol{R}$ satisfies $\boldsymbol{Q}$; (ii) it is causally or metaphysically necessary that, if $x\ R_2$s something and $x\ R_3$s, then $x\ R_2$s the proposition that $x\ T$s something; and (iii) $x\ R_2$s q.

In fact, however, the intensionality worry is unfounded. True, *singular terms* occurring in intensional contexts cannot be existentially generalized. For example, even if '*x* thinks that the smartest spy is a spy' is true, its existential generalization 'For some *y*, *x* thinks that *y* is a spy' might well be false. The intensionality worry is that the argument overlooks this familiar point. But there is an important difference between our main argument and the above example involving 'the smartest spy.' Specifically, when we Ramsified in our argument, we existentially generalized, not on a singular term ('the smartest spy'), but rather on an embedded *predicate* ('thinks'). This existential generalization is on a par with the following inference: *x* thinks that he or she hurts; therefore, for some *R*, *x* thinks that he or she, *R*s. This is valid in the logical settings in which functionalists themselves intend to Ramsify.

Look at the matter this way. The proposition that is the semantic value of 'he thinks something' in the antecedent of Self-intimation Principle is the same as the proposition that is the semantic value of the 'that'-clause 'that he thinks something' in the consequent. In each case, it is the proposition *that he thinks something*. This is why the following truism holds: 'he thinks something' means *that he thinks something.* The correct logical analysis of this proposition is given in terms of the thinking relation, the relation (i.e., intension) expressed by the psychological verb 'thinks.' Since functionalists intend their predicate variables to replace psychological verbs and since each occurrence of 'thinks' is semantically correlated with one and the same relation (intension), functionalists would have us replace each occurrence with one and the same predicate variable 'R_2.' This is what we did in the argument. So our argument goes through and involves no equivocations over intensionality.[12]

The entire issue of intensionality does not even arise in the case of the psychological relation of self-attribution (which Lewis (1979) and Chisholm (1981) focus upon). On analogy with the Self-intimation Principle, the following

But this is not the sort of definition allowed by functionalism, for the intended value of its quantified predicate variable '*T*' is the thinking relation itself, not a first-order realization of it. Instead, this definition is a special case of what I will call a nonreductive functional definition (see sections 4–5).

[12] Some people have tried to avoid the Self-consciousness Argument by proposing an intensional logic that deals with both properties (attributes) and concepts. In such a setting one of two things happens to the original Self-consciousness Argument. Either a version of that argument still goes through but is somewhat more complicated (see Bealer 1997: 83). Or else the new Ramsified definitions turn out to be nonreductive (in the sense isolated in section 4). In the latter case, those definitions might well be correct, but if they are, functionalism's solution to the Mind–Body Problem collapses in the manner discussed in section 5. For now, suffice it to say that invoking concepts does not help to avoid the problem in the case of the psychological relation of self-attribution (cf. the next paragraph in the text). For it is undeniable that ordinary people attribute to themselves the *attribute* of being in pain. The relation of attribution clearly relates people to properties (attributes) not concepts. So the self-consciousness argument applies just as it did before; the more complicated apparatus of both concepts and properties (attributes) does nothing to prevent it.

principle partially characterizes the self-attribution relation: if *x* has the property of being in pain and *x* has the property of introspecting, then *x* will self-attribute the property of being in pain. Here the occurrences of 'the property of being in pain' following 'has' in the antecedent and following 'self-attributes' in the consequent are both plainly extensional. So, uncontroversially, the Ramsification of this principle is: if *x* has R_1 and *x* has R_3, then *x* is related by R_2 to R_1. Then the rest of the self-consciousness argument goes through *mutatis mutandis*. The absurd conclusion follows, namely, that it is commonplace for ordinary persons *x* to attribute to themselves first-order realizations of the property of being in pain (say, the property of having firing C-fibers), rather than the property of being in pain itself.[13] For this reason, the standard style of Ramsified definition does not work generally: it fails in the case of the psychological relation of self-attribution.[14]

To avoid our argument, what functionalists need is a way of blocking the quantification of embedded mental predicates ('thinks,' etc.) in psychological theory *A*, and they must do this in a way that does not leave them with undefined psychological expressions on the right-hand sides of their definitions. This can be accomplished by treating embedded predicates as standing for mere syntactic entities—mere linguistic *representations*. Language-of-thought functionalism is designed to do just this. But it does not solve the underlying problem; it only hides it.

3. LANGUAGE OF THOUGHT

Language-of-thought functionalism resembles a common two-step technique of giving functional definitions. According to this technique, 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 state of thinking, a state of desiring, etc. Following that, one attempts to define what it is for a mental state of a given type to have *p* as its content; that is, to define what it is for a state of thinking to have *p* as its content, what it is for a state of desiring to have *q* as its content, etc. Putting the two steps together, one then obtains fully general definitions of what it is for *x*

[13] With this simpler argument in mind, it is easy to see that a wholly analogous argument shows that it would also be commonplace for ordinary people to self-attribute first-order realizations of the property of thinking something (rather than the property of thinking something itself).

[14] Against this argument it might be objected that self-attribution is a nonbasic mental relation which is to be defined, not by means of Ramsification, but rather *directly* in terms of the thinking relation itself. (This approach is not available to functionalists, such as Lewis, who take self-attribution to be definitionally prior to thinking.) On this approach, self-attribution might be defined as follows: *x* self-attributes *F* iff_{def} *x* thinks that he or she is *F*. But this just concedes the larger point. For, uncontroversially, this embedded occurrence of '*F*' on the right-hand side of the definition is externally quantifiable. So the Self-consciousness Argument goes through in its original form.

to think p, desire q, etc. For example, x thinks p iff_{def} x is in a state of thinking and that state has p as its content.[15]

Language-of-thought functionalism resembles this approach except that, in step one, tokenings of Mentalese sentences in modules (metaphorically, a Thinking Box, Desiring Box, etc.) take the place of mental state types and, in step two, content is assigned to the Mentalese sentences that might be so tokened.[16] Thus, x thinks p iff_{def} there is a Mentalese sentence s tokened in x's Thinking Box and s's content is p; x desires p iff_{def} there is a Mentalese sentence s tokened in x's Desiring Box and s's content is p; and so forth.

The success of step one is incompatible with the metaphysical possibility of a nonphysical, purely mental being.[17] The problem this possibility would create for language-of-thought functionalism is that, for purely mental beings, there could be no physical medium in which the requisite Mentalese sentences could be tokened and no physical modules in which to house these tokens.[18] Accordingly, general definitions of the standard mental relations (thinking, desiring, etc.) would be out of the reach of language-of-thought functionalism.[19] Of course, many language-of-thought functionalists reject the possibility of a purely mental being, so the argument of this section will not turn on it. The reason for mentioning it here is that, as we will see, language-of-thought functional definitions stand no chance of being correct unless they are

[15] One instance of this method is just an elaborate form of Ramsification. For example, a state t would be a state of thinking iff_{def} there exist first-order realizations $\boldsymbol{R}$ satisfying $\boldsymbol{A}$ and $t =$ the state of being related by R_2 to something. And a state t of thinking would have p as its content iff_{def} there exist first-order realizations $\boldsymbol{R}$ satisfying $\boldsymbol{A}$ and $t =$ the state of being related by R_2 to p. Of course, such Ramsified formulations of the two-step approach are plainly subject to the style of argument given above.

[16] See, for example, Fodor (1987). Incidentally, at certain points in what follows I take the liberty of using ordinary quotation marks where Quinean corner quotation marks are strictly speaking called for.

[17] Not only do many Ramsifying functionalists accept this possibility, but so do various language-of-thought functionalists. For example, Fodor (1981) recognizes "the possibility that systems as diverse as human beings, calculating machines, and disembodied spirits could all have mental states." Furthermore, there are serious *arguments* supporting this possibility: for example, Yablo's (1990) reformulation of the traditional conceivability argument, and my own (1994) reformulation of the traditional certainty argument. Fodor's acceptance of this possibility creates a tension within his overall view.

[18] "Non-physical stuff" is an oxymoron, a metaphysical impossibility: "ectoplasm" and its ilk are just silly—no serious philosopher subscribes to such things. Moreover, the ectoplasm move would not avoid the problem unless it is assumed that some such medium would be necessary for the existence of a purely mental being. But there is no reason to accept this assumption. Indeed, most philosophers who accept the possibility of purely mental beings think they would have to be *simple* beings—not "made" or "composed" of anything.

[19] Here, then, is an advantage of Ramsifying functionalism: it is at least prima facie consistent with this possibility—since, in the case of, say, angels, the requisite first-order realization could be the relation of angel-thinking. Given that language-of-thought functionalism is simply incompatible with the possibility of a purely mental being, it bears a certain *ontological* resemblance to explicitly materialist formulations of Ramsifying functionalism, for example, those in which the first-order realizations are *explicitly* restricted to first-order physical properties and relations.

reformulated as *nonreductive* functional definitions (in the sense of section 4). In this case, however, language-of-thought turns out to be an inessential third wheel.

It is in step two that the problem of self-conscious thought resurfaces. Assume that the content-of relation is somehow defined for Mentalese non-psychological expressions.[20] The question is whether it can be defined for Mentalese *psychological* predicates, specifically, Mentalese predicates ('*T*,' '*D*,' etc.) for the standard mental relations (thinking, desiring, etc.). For, unless this can be done, one will not have defined any of these relations. But the familiar definitional strategies lead to vicious circularity (and other failings)—unless, of course, we return to Ramsified definitions and, with them, some version of our original problem.

The vicious circle is immediately evident in the following candidate definition: the content of '*T*' $=_{\text{def}}$ the relation of thinking. For the contemplated definition of the relation of thinking (*x* thinks *p* iff$_{\text{def}}$ there is a Mentalese sentence *s* tokened in *x*'s Thinking Box and *s*'s content is *p*) requires that we have already specified the content of '*T*'. The same problem besets the following definition: the content of '*T*' $=_{\text{def}}$ the relation holding between *x* and *q* such that a Mentalese sentence *s* whose content is *q* is tokened in *x*'s Thinking Box. But this too is circular, for the content-of relation for arbitrary Mentalese sentences *s* is invoked on the right-hand side: in order to define this relation, one must first define the content-of relation for the primitive predicates that can occur in those sentences *s*—including, in particular, the predicate '*T*' for the relation of thinking.

Another strategy for avoiding the circularity problem would be to adopt a causal account of content. For example, let '*C*' be a Mentalese predicate for one or another macroscopic physical property. Then a (highly oversimplified) causal account of its content might go as follows: the content of '*C*' $=_{\text{def}}$ the property *F* such that in normal conditions there being an *F* in the presence of a subject causes '$(\exists x)Cx$' to be tokened in the subject's Thinking Box. (For example, on the assumption that in normal conditions there being a cow in the presence of the subject causes '$(\exists x)Cx$' to be tokened in the subject's Thinking Box—and being a cow is in normal conditions the only macroscopic physical property playing this causal role—the property of being a cow would be the content of '*C*'.) The corresponding causal account of the contents of Mentalese psychological predicates (e.g., '*T*') would then be something like this: the content of '*T*' $=_{\text{def}}$ the relation *R* such that, for any *q*, if the subject is *R*-ing *q* in normal conditions, the subject's *R*-ing *q* causes '*i T s*' to be tokened in the subject's Thinking Box, where s is some Mentalese sentence whose content was previously defined to be *q*. There are many problems with this approach. I will mention two—each

[20] Whether this is feasible is open to serious doubts. One threat comes from the possibility of deviant Quinean interpretations of Mentalese. See Bealer (1984).

of which shows that this approach is incompatible with language-of-thought functionalism and, hence, may not be used to solve its circularity problem.

First, the indicated style of causal account would be correct only if the relation *R* (i.e., the content of '*T*') were the thinking relation itself. But the account must be incorporated into the language-of-thought definition of the thinking relation itself (stated two paragraphs above). So this definition of the thinking relation quantifies over the very relation being defined and, hence, accords to the thinking relation an ontological primacy inconsistent with the primary tenet of functionalism (that mental relations can be defined wholly in terms of the PATTERN of their *ontologically prior* first-order realizations). Indeed, on this account, the causal status of certain events involving the thinking relation is like that of certain events involving physical properties and relations—e.g., the property of being a cow, the relation of being tokened-in—inasmuch as all these events have the power to cause tokenings of relevant Mentalese sentences. The proposed language-of-thought definition would thus qualify as a thoroughgoing non-reductive functional definition, in the sense of section 4 (except that it gratuitously builds in the paraphernalia of language-of-thought).

The second problem with this style of causal account of the content of '*T*' is that it explicitly requires *mental-to-physical* causation—specifically, the mental event of the subject's thinking *q* must cause a certain physical event, namely, the tokening of '*i T s*' in the subject's Thinking Box. But such mental-to-physical causation violates the causal picture on which language-of-thought functionalism is founded. On that picture, what causes '*i T s*' to be tokened in a subject's Thinking Box is a physical language-of-thought event, such as the event of *s*'s being tokened in the subject's Thinking Box (where *s* is a language-of-thought sentence whose content is *q*).[21] In other words, the property that would be causally relevant to the tokening of '*i T s*' is not the intentional property of thinking *q* but rather a physical realizer property.

Let me spell this out. Consider the array of law-governed transitions from mental event to mental event (or mental state to mental states), and consider the corresponding transitions from language-of-thought event to language-of-thought event. The basic language-of-thought causal picture has it that the latter array is not just law-governed but is founded upon genuine physical causal relations holding among these language-of-thought events; in the idiom of causal relevance, it is the associated physical language-of-thought realizer properties (not the intentional properties corresponding to them) that figure in the causal

[21] I just spoke as though there is a difference between the tokening of *s* and the subject's thinking *q*; this is so on a fine-grained view of events. On certain coarse-grained views of events, there is no such difference; but then the question just shifts to which property is causally relevant, the property of thinking *q* or a correlated but distinct language-of-thought realizer property. In the text, I employ both idioms to help make clear that no question is being begged.

explanation of these physical-to-physical transitions. For example, suppose the following reports a lawful transition: if a subject is thinking *q* and is engaging in introspection, he will think that he is thinking *q*; and suppose that in a given subject's language-of-thought *s* is a sentence whose content is *q*. Then, on the language-of-thought causal picture, not only would it be nomologically necessary that a tokening of *s* in the subject's Thinking Box is followed by a tokening of '*i T s*', but in the envisaged situation a situation the tokening of *s* would be the cause of the tokening. The language-of-thought realizer property (having *s* so-tokened) is the property that is causally relevant to the tokening of '*i T s*'; the intentional property (thinking *q*) is causally irrelevant (contrary to the proposed account of the content of '*T*').[22]

(Here is Fodor commenting on this basic language-of-thought causal picture (1987, 140): "[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." So, for example, Fodor would hold that the intentional property of thinking q does not figure in any psychological mechanism and thus does not figure causally in producing tokenings of '*i T s*'; instead, the language-of-thought realizer property is what is causally relevant. He would hold, moreover, that this physical realizer property is causally relevant to subject's thinking that he is thinking *q* whereas the mental property of thinking *q* is not. And this generalizes: for Fodor, the causally relevant properties are always physical realizer properties, never intentional properties. Fodor's view is thus a form of epiphenomenalism. By contrast, the sort of nonreductive functional definitions suggested in §4 open up the possibility of an account of mental causation (see Bealer, 2007) that avoids epiphenomenalism and preserves most of our commonsense beliefs about mental causation; this is a further count in favor of this nonreductive functionalism and against Fodorian and other epiphenomenalist functionalisms.)

If these and similar problems block a causal account of the content of the Mentalese psychological predicates ('*T*,' '*D*,' etc.), we are still left with the circularity problem. What alternatives are there? A common technical proposal is to resort to a Tarski-like hierarchy of thinking relations—thinking$_0$,

[22] The reason, as I understand it, that language-of-thought functionalists are committed to this causal picture is that it is needed—or at least they believe it is needed—to explain what it is for a system to be a physical realization (or physical implementation) of a psychological system: if causation were not invoked in the indicated way, all manner of physical systems would wrongly qualify as physical realizations (thus undermining the language-of-thought definitions of mental properties). See, e.g., the Appendix in Fodor (1987). David Chalmers (1996b) also provides a nice explanation of the role causation plays for functionalism (and, presumably, he would enlist causation to play the same role in his own functionalist account of intentional properties).

thinking$_1$, . . .—defined as follows: x thinks$_0$ q iff$_{\text{def}}$ some nonpsychological Mentalese sentence whose content is q is tokened in x's Thinking Box. The content of 'T_0' $=_{\text{def}}$ thinking$_0$. Next x thinks$_1$ q iff$_{\text{def}}$ some level 1 Mentalese sentence whose content is q is tokened in x's Thinking Box. The content of 'T_1' $=_{\text{def}}$ thinking$_1$. And so on.

But it is now widely recognized that such hierarchy approaches lead to a distorted treatment of our actual psychological attitudes (just as Tarski's hierarchy approach leads to a distorted theory of truth; see Kripke 1975). For example, you might say, "Most people think many things." And I might reply, "I certainly do; in fact, you have just asserted one of them." In this little dialogue, you assert a proposition involving the relation of thinking, namely, the proposition that most people think many things. I reply by affirming a certain *instance* of the proposition you asserted, namely, that *I* think many things. Then I go on to provide an example of one of the things to which I stand in the thinking relation, namely, the original proposition you asserted (that most people think many things). This, however, is a proposition involving the very relation of thinking just invoked. If this were not so, my *anaphoric* use of 'one of them' would make no sense. Examples like this one are not at all exceptional; they typify our everyday thought and discourse about cognition. Much the same point is tellingly illustrated by Descartes' *Cogito*. Suppose I think that I am thinking something. (In symbols, i Think $[(\exists q)$ i Think $q]$.) The proposition to which I stand in the thinking relation involves that very relation of thinking. And this proposition is made true just by the fact that I am standing in that very relation to it. This is the point of the *Cogito*—and what is compelling about it. The moral is that, like truth, thinking and other mental relations are *type-free*. The proposed hierarchy picture belies this fundamental fact.

Perhaps, however, the hierarchy picture is still useful theoretically, allowing one to *construct* the thinking relation from the hypothesized relations thinking$_0$, thinking$_1$, etc. The most common proposal is to identify the thinking relation with the *union* of these hypothesized relations. To see why this fails, notice that the thinking relation is distinct from each of the hypothesized relations. For the range of the thinking relation includes such psychological propositions as the proposition that someone thinks something, whereas the range of thinking$_0$ includes, by definition, only nonpsychological propositions. Similarly, the range of thinking$_1$ does not include the proposition that someone thinks something; rather it includes such propositions as the proposition that someone thinks$_0$ something. But these two propositions are distinct, for thinking $\neq$ thinking$_0$, as we have just seen. The argument generalizes in the obvious way. It follows that the proposition that someone thinks something does not belong to the range of any thinking$_n$ relation and so does not belong to the range of the union of the thinking$_n$ relations. Indeed, *not one* proposition involving the thinking relation belongs to the range of the union of the thinking$_n$ relations. But countless

such propositions belong to the range of the thinking relation itself. Hence, the thinking relation and the union are *very* different indeed.[23]

The problem with the hierarchy approach is that we can have attitudes toward type-free general propositions that are about nothing other than those very attitudes. This problem also spells defeat for the other standard way of trying to approach mental relations in stages—namely, the standard *recursive* approach. Unlike the hierarchy approach, this approach supposes that there is a single relation of thinking, desiring, etc. It aims to define the content-of relation for Mentalese in inductive stages. At the initial stage, we define the content-of relation for all Mentalese nonpsychological primitives. Then there are two sorts of inductive clauses. One defines the content-of relation for the various categories of complex expressions (existential generalizations, negations, etc.). The other defines the content-of relation for Mentalese psychological primitives ('*T*,' '*D*,' etc.): the content of '*T*' $=_{\text{def}}$ the relation holding between *x* and *q* such that a Mentalese sentence whose content is *q* is tokened in *x*'s Thinking Box. And so forth. Superficially, this has the form of an acceptable inductive definition. Our previous discussion, however, reveals what is wrong with it. Type-free Mentalese sentences such as '$(\exists q)\ i\ T\ q$' can be tokened in one's Thinking Box. Accordingly, the above inductive clause fixes the content of '*T*' only if the content of '$(\exists q)\ i\ T\ q$' is fixed earlier in the induction. But the content of '$(\exists q)\ i\ T\ q$' is fixed only if the content of '*T*' is fixed still earlier in the induction. The inductive clauses thus fail to fix any of these contents.[24]

Once again, the problem is that we can have attitudes toward type-free general propositions that are about nothing other than those very attitudes. They cannot be built up in stages: whether you are thinking that you are thinking something is in principle independent of the other things, if any, you might be thinking; relative to them, it is a primitive fact. This leaves us where we were before the detour into stage-wise approaches, either with a viciously circular definition of the content-of relation or with no definition at all.[25]

[23] Thus, the following proposal also fails: *x* thinks *q* iff for some *n*, *x* thinks$_n$ *q*. Of course, *within* standard type theories one cannot even quantify over levels *n*.

Note also that it is possible for me to be thinking that I am thinking something and to be thinking no proposition involving any thinking$_n$ relation. Conversely, for any *n*, it is possible for me to be thinking that I am thinking$_n$ something and not be thinking that I am thinking something. These things are possible because, as noted earlier, thinking can be highly *focused* in its propositional objects: at a given moment a person can be thinking a specific proposition *q* and not be thinking any other nearby *q*. Indeed, someone skilled in meditation arguably can be consciously and explicitly thinking that he or she is thinking something whether or not he or she is at the moment consciously and explicitly thinking anything else.

[24] This difficulty can be overcome by means of a diagonal construction, but the result is just a complicated variant of the sort of nonreductive functional definition considered in the next section and so is subject to the same conclusions.

[25] This discussion also shows that infinitary disjunctive definitions of the standard mental relations fail because they are circular: for example, *x* thinks *q* iff$_{\text{def}}$ (*x* is in physical state *S* and *q* = the proposition that something is a horse) . . . or (*x* is in physical state *S*′ and *q* = the

There are, of course, alternative language-of-thought proposals. But I know of none that helps to solve the underlying problem of self-conscious thought.[26]

4. NON-REDUCTIVE FUNCTIONAL DEFINITIONS

At this point, there is a very natural response to the problem of self-conscious thought, namely, returning to the original functional definitions but this time suppressing the problematic invocation of realizations and focusing instead on the mental properties themselves. According to the original functional definitions, to be in pain is to have some *first-order realization* of the property of being in pain; to think q is to be related to q by some *first-order realization* of the thinking relation; and so on. The idea is to expunge these restricted quantifications over first-order realizations and to replace them with quantifications that admit as values mental properties themselves. The result would be definitions like the following: to be in pain is to have some property that plays the pain-role in $\boldsymbol{A}$; to think q is to be related to q by some relation that plays the thinking-role in $\boldsymbol{A}$. More formally,

x is in pain iff_{def} there exist properties $\boldsymbol{R}$ satisfying $\boldsymbol{A}$ and x has R_1.

x thinks q iff_{def} there exist properties $\boldsymbol{R}$ satisfying $\boldsymbol{A}$ and x is related by R_2 to q.

The crucial difference is that the property expressed by the entire right-hand side is among the values of the variables ('R_1,' 'R_2,' etc.) occurring within the right-hand side. Thus, unlike the original functional definitions, which were formulated in the logical setting of a *predicative type theory* (Putnam, 1970, is very clear on this point), these definitions are formulated in an *impredicative type-free* logical setting. So, if these definitions are correct, the property of being in pain is itself among the properties over which the first definition quantifies; likewise, the thinking relation is itself among the relations over which the second definition quantifies.[27]

Functional definitions of this sort have a significant feature. Since the standard mental properties and relations are themselves satisfiers of $\boldsymbol{A}$, they are being defined in terms of their interaction with *themselves and one another*. In view

proposition that someone thinks something) or . . . or (x is in physical state S'' and $q =$ the proposition that thinking is a mental relation . . .).

[26] For example, one might attempt a language-of-thought variant on the kind of Ramsified functional definitions discussed in section 2, but then the argument given there can be repeated to show that the content of 'T' would be a first-order realization rather than the thinking relation itself. Alternatively, one might attempt a language-of-thought variant on the kind of nonreductive functional definitions discussed in the next section, but the resulting definitions would be inconsistent with the ontological picture upon which functionalism is based.

[27] Sydney Shoemaker's (2001) response to the Self-consciousness Argument is to reject standard Ramsified functional definitions, which he previously (1981) espoused, and to adopt instead the style of nonreductive functional definitions proposed here and in Bealer (1997).

of the problem of self-conscious thought, this feature is unavoidable in any successful functional definition, for it is this feature that opens up the possibility of getting right the contents of our everyday self-conscious thoughts. The price of this benefit, however, is that these definitions abandon the primary tenet of functionalism, namely, that the standard mental properties be definable wholly in terms of the PATTERN of *ontologically prior realizations*. These definitions are therefore *nonreductive* in the sense that, unlike the original functional definitions, they do not equate mental properties with (second-order) constructions from ontologically prior realizations.[28] On the contrary, these definitions endow mental properties with an ontological primacy inconsistent with the standard functionalist picture: mental properties are now taken to be antecedently given ontological primitives (specifically, first-order irreducibly mental properties) already there waiting to constitute the content of our thought. These definitions merely locate mental properties within the space of ontologically primitive properties.

As indicated, the reason these nonreductive definitions are more promising is that the standard mental properties are admitted as satisfiers of *A* and (unlike the original functional definitions) they do not *require* that properties besides the standard mental properties (e.g., associated physical realizations) satisfy *A*. But this does not by itself guarantee that these definitions are successful. For it does not rule out the existence of unwanted satisfiers of *A* (e.g., physical realizations). To guard against this possibility, our nonreductive definitions might be strengthened in various ways. For example, the initial string of predicate quantifiers in the definitions might be explicitly restricted to properties that are not physical realizations and that are "natural" (i.e., not ad hoc Cambridge properties). In addition, *A* itself might be strengthened in various ways. For example, *A* might be strengthened *modally*—for instance, by requiring that its satisfiers satisfy it *necessarily* and by requiring that it be *possible* for its individual clauses to be satisfied nonvacuously. When *A* is strengthened in this and perhaps other ways, it is plausible that *A* would *implicitly define* the standard mental properties—that is, *A* would be *uniquely* satisfied by the standard mental properties. For many people, the resulting definitions are the ideal—what one should strive for in a good functional definition.

Now since successful functional definitions of some sort are a precondition of functionalism's solution to the Mind–Body Problem and since the problem of self-conscious thought evidently forces functionalists to accept some sort of nonreductive definitions, we may assume for the purpose of the remaining discussion that nonreductive functional definitions of some sort are successful. To simplify this discussion, it will be convenient to assume that the simple style of nonreductive definitions given at the outset of the section are correct.

[28] The same thing holds true of more complex nonreductive definitions—for example, definitions like those considered in sections 2–3 and accompanying footnotes.

This simplifying assumption is harmless, for it will be evident that each of our remaining points would hold even if some more complicated nonreductive definition were needed (cf., notes 17 and 19).

5. CONSEQUENCES

The adoption of nonreductive functional definitions, however, has a major consequence. Before coming to it, let us take stock.

Mental relations have dual roles. They relate subjects to propositions, and they are commonly contents of those propositions. One and the same relation is involved twice over when a person thinks that he or she is thinking something. Our original argument against functionalism turned on such phenomena: in the original functional definitions, both the embedded and unembedded occurrences of a psychological predicate were quantified with one and the same quantified predicate variable; accordingly, those definitions wrongly implied that propositions involving first-order realizations would be typical objects of thought. Language-of-thought functionalism seemed to avoid these dual quantifications by treating embedded psychological predicates as standing only for representations rather than for mental properties and relations themselves. But this only hid the problem, for then the contents of Mentalese psychological predicates needed to be identified. It turned out that this could not be done without circularity (and other problems). The error of functionalism as traditionally formulated was to think that mental properties and relations are in one way or another *constructible* from ontologically prior realizations. They are not. Functionalists evidently have no alternative but to adopt nonreductive functional definitions, thereby abandoning their primary tenet (that mental properties are definable wholly in terms of ontologically prior realizations). The phenomenon of self-conscious thought teaches us that mental properties must be antecedently given ontological primitives already there waiting to constitute the content of our thought; they must be part of the primitive make-up of the world. Indeed, by virtue of their primitive self-reflexive loops, mental relations might well stand as our very paradigm of irreducibility.

As I noted at the outset, perhaps functionalism's major conceptual attraction for cognitive science was that it promised a materialistically acceptable solution to the Mind–Body Problem. The idea was to employ reductive functional definitions to explain the relationship between physical and mental properties. Specifically, when such definitions are substituted into the following, the result was supposed to be a tautology: if a thing has some system or other of first-order physical properties that fit the PATTERN, then it has the associated mental properties. In this sense, the body–mind relationship would be completely transparent and unmysterious—just a matter of logic and definitions. We have seen, however, that the envisaged functional definitions fail and that functionalists evidently

have no choice but to adopt some form of nonreductive functional definition. When this is done, however, the envisaged definitional tie between physical and mental properties is broken. No matter what constellation of first-order physical properties ***F*** you might have (where ***F*** may somehow encode physical laws if you wish), it is not a matter of logic and definition that, if you have ***F***, you also have the mental properties that you in fact have.[29] From the point of view of pure logic, your first-order physical properties tell us absolutely nothing whatsoever about whether, in addition, you have those distinct, ontologically primitive properties that are defined by the nonreductive functional definitions. On the contrary, the body–mind relationship is a primitive nonlogical relationship. Thus, functionalism's solution to the Mind–Body Problem fails.

To be sure, the highly substantive scientific thesis that (in the actual world) mental properties have only physical realizers is compatible with the foregoing critique. But this (contingent) scientific thesis does not, on its own, illuminate the nature of the body–mind relationship. What accounts of this relationship are there? One view is that it is a nomological relationship: the discovered correlation between physical and mental properties holds as a (contingent) law of nature. Evidently, the only alternative to the nomological view is that the body–mind relationship is a mysterious kind of brute metaphysical necessity, defying any further explanation. But this alternative is unacceptable, given that the nomological account is available. To begin with, the hypothesis of brute metaphysical necessities has no more intuitive support than the nomological hypothesis (on the contrary, it is the other way around, as the much-discussed body of anti-materialist intuitions attests). Nor does the metaphysical hypothesis have any explanatory advantages over the nomological hypothesis; both hypotheses have exactly the same empirical consequences. In addition, the nomological hypothesis is distinctly more economical: nomological necessities need to hold only in a local sphere of worlds whereas metaphysical necessities have to hold in all possible worlds. Viewed another way, the hypothesis of brute metaphysical necessities takes a strong stand in modal metaphysics when none is required. And, what is more, accepting such hypotheses in situations like this is clearly incompatible with standard practice in modal metaphysics.

If this is correct, scientifically minded philosophers and cognitive scientists have no rational alternative but to accept the nomomogical account of the body–mind relationship. Of course, this choice in no way impedes the pursuit of a scientific understanding of the mind. Discovering a law of nature is no less scientific than discovering a property identity (as identity theorists hoped to do)

[29] Specifically, it can be shown that, given the nonreductive definitions, the standard sort of logical inference route from ***F*** to your mental properties is unavailable. Moreover, models can be constructed to show that this relationship cannot be a less direct sort of logical relationship. Such models also show that the body–mind relationship fails to hold as a matter of logic and definition even if one accepts some more complicated style of nonreductive functional definition (such as those considered at various points earlier in the paper). Bealer (1999) develops these points in detail.

or discovering that a certain relationship holds as a matter of logic and definition (as functionalists hoped to do). So on this path nothing scientific is lost; indeed, the project of cognitive science is entirely clear: find the laws. The only casualty is materialism, a metaphysical doctrine to which science was never committed in the first place.