MIDWEST STUDIES IN PHILOSOPHY, IX (1984)

Mind and Anti-Mind: Why Thinking Has No Functional Definition

GEORGE BEALER

Functionalism is perhaps the most prominent theory of mind today. The central thesis of functionalism is that the standard mental relations (or properties or states) are uniquely determined by their causal roles in functioning organisms. That is, the principles of psychology specify the characteristic way that (behavioral or physiological) input, the standard mental relations such as belief and desire, and (behavioral or physiological) output are causally arranged; and the central idea of functionalism is that, e.g., belief's characteristic causal role can be fulfilled by exactly one relation–namely, belief itself. Clearly, then, the most direct way to refute functionalism would be to show that there are relations that demonstrably differ from the standard mental relations and that, nevertheless, could fulfill the same causal role as those mental relations.

However, it is unsatisfactory to leave the discussion couched in these imprecise terms: imprecise positions are difficult to defend or to refute. What exactly are causal roles? What exactly are their identity conditions? What sorts of things have them? Is it mental types or tokens that are uniquely determined? It is for good reason, therefore, that the imprecise functionalist thesis is often reformulated more precisely as follows: the principles of psychology, taken together, implicitly define the standard mental relations. That is, these relations–and no others–make the principles of psychology true when we hold constant the interpretation of the physical and logical constants contained in these principles. There are well-known techniques for converting implicit definitions into direct definitions,[2] and thus many functionalists also put their central thesis as follows: the standard mental relations have direct functional definitions based on the principles of psychology.[3] So when functionalism is formulated in one of these precise ways, the most direct way to refute it is to show that, in addition to the standard mental relations, there exists a demonstrably different system of deviant relations that make the principles of psychology come out true when we hold constant the interpretation of their physical and logical constants. The purpose of this paper is to construct

just such a system of deviant relations. This new system of relations in effect comprises an "anti-mind"–a system causally and functionally indistinguishable from the system of familiar relations that comprises the true mind. The existence of anti-mind simultaneously refutes all versions of functionalism whether behavioral, physiological, or computer-theoretic in emphasis.

What makes the construction possible is, in a word, *intensionality*, specifically, those "fine-grained" distinctions that can exist among our intentional states. There are some well-known arguments showing that no elementary behavioral or physiological reduction can capture this aspect of those states. A generalized version of these arguments provides the key to the construction of anti-mind.[4]

THE NECESSITY FOR A GLOBAL REFUTATION

To understand more clearly the conditions that a refutation of functionalism must meet, let us begin with a brief survey of the currently popular criticisms of functionalism.[5] Although these criticisms abound, I believe that either they miss our target–namely functionalism's central thesis–or they are inconclusive. I will briefly sketch why I believe this. To minimize controversy, I will try to avoid taking issue with the assumptions upon which a given criticism is based.

Consider first the criticism given by some eliminative materialists.[6] According to this criticism, functionalism is just unscientific: with the progress of science, the standard mental notions will join the company of the unscientific notions belonging to such theories as alchemy and astrology. However, this criticism has no bearing on the thesis that our mental notions have functional definitions. Even if mental notions are destined to play no role in the science of the future, nevertheless, the functionalist's definitions of mental notions could be logically correct. A functional definition could specify logically necessary and sufficient conditions for belief or desire, for example, regardless of whether these notions are needed in the prediction or explanation of our behavior. Indeed, the correctness of such a definition is a *modal* fact which holds independently of whether any extant beings happen to have beliefs or desires. Furthermore, philosophers will continue to be interested in the correctness of such definitions regardless of the course of future science. For it is one of the jobs of philosophy to determine what it *would* take for a being to have beliefs or desires. And surely it is logically possible for some being to have a belief or a desire.

The remaining criticisms may all be viewed as counterexamples to the kind of definitions posited by functionalists. The first one derives from recent attacks on individualism.[7] Individualism is the doctrine that the identity of a person's mental states depends exclusively upon the "internal" state of the individual at the moment; social factors are entirely incidental. However, according to the recent anti-individualist doctrine, two individuals who are in qualitatively the same internal states but different social situations can be in different mental states. One of the arguments goes as follows. Suppose that we have doppelgängers on "Twin Earth" who use the term 'water' to refer to a water-like stuff XYZ on occasions otherwise just like those on which we refer to the stuff water. Suppose that a linguistically competent Earthling and his doppelgänger are in qualitatively the same internal states, and suppose that each of them sincerely utters the sentence 'Water is wet'. Then, the critic asserts, the Earthling would believe that water is wet but the Twin Earthling would believe that XYZ is wet. If the critic is right, this would be a counterexample to individualism. And to the extent that functionalism is committed to individualism, it would be a counterexample to functionalism. However, even if this attack on individualism is right, it does not clearly damage functionalism. First, the functionalist can deny that he is committed to individualism. He might relax his criteria for what counts as an "input condition" upon an individual; specifically, he might count as an input condition any physical feature in the individual's extended environment. Since water is found in the Earthling's extended environment where XYX is found in that of his doppelgänger, their input conditions differ, and this suffices to explain why their mental states differ. Second, the functionalist is free to invoke a more sophisticated theory of our psychological make-up. According to this theory, in the psychological make-up of every person there is an "individualist core," i.e., a body of mental states that conform to all the traditional individualist doctrines and that are immune to all the anti-individualist arguments.[8] (A person's purely qualitative self-presenting states are one example of the sort of state typically belonging to his individualist core.) Although the identity of some mental states may be determined *in part* by social factors, the identity of a mental state is never determined *exclusively* by social factors; the mental states in the persons's individualist core are always a factor in determining the identity of his non-individualistic mental states. If some such theory of psychological make-up is right, then functionalists can circumvent the anti-individualist attack by restricting functionalism to mental states of the type belonging to individualist cores. Accordingly, the functionalist's central thesis would be that all mental states of this basic type are uniquely determined by their causal roles in functioning organisms. Though more restricted, this thesis still would be very important philosophically.

The next type of counterexample is that of a system that is in significant respects functionally isomorphic to a human being and yet does not strike us as having genuine mental states. Counterexamples of this type range from everyday digital computers to Ned Block's China example, which goes as follows.[9] The number of people in China is about the same as the number of cells in a human brain. Block suggests that the Chinese people could be so coordinated that China, taken as a whole, could be functionally isomorphic to a human brain. He holds, however, that China would nevertheless have no genuine mental states. Unfortunately, examples of this type are inconclusive. To begin with, the functionalist can challenge Block's assumption that China's physical states could ever possess physical causal roles isomorphic to those possessed by the states of a functioning person's brain. Second, even if China's *physical* states could be like this, more is needed for a genuine counterexample: China must also be in states possessing causal roles isomorphic to those possessed by the *mental* states of the per-

son whose brain it is. But what guarantees that this second isomorphism would hold? On the one hand, since not all versions of functionalism are committed to the mind/body identity theory, Block is not free to assume that the second isomorphism follows directly (by Leibniz's law) from the supposed physical isomorphism. On the other hand, perhaps Block would say that the second isomorphism follows from the physical isomorphism indirectly via some special new kind of supervenience principle. However, it is hard to think of any credible supervenience principle that would do the job. After all, China and the person whose brain we are considering are in very different physical states. Finally, if this gap in Block's reasoning is not enough to undermine his example, the functionalist can always just assert that entities like China would be in genuine mental states; for example, Douglas Hofstadter holds, ". . . there could be a 'big person' built out of ordinary people–but probably it would take many trillions of people, not just hundreds of millions, on the assumption that one ordinary person is simulating the function of a single cell in the giant person."[10] Maybe the functionalist would be wrong, but how is the critic to *show* that he is? We need a counterexample that is *demonstrably* immune to this easy dodge on the part of the functionalist.

Another candidate counterexample, which is far more threatening to functionalism, derives from the prospect of an inverted spectrum.[11] The following is a forceful way to formulate this counterexample.[12] Let a function f map the shades from one end of the color spectrum onto shades from the other end and conversely. Consider the relation sensing*: x senses* y iff$_{df}$ x *senses* $f(y)$ if y is a shade and x senses y if y is not a shade. The worry is that the relation sensing* satisfies all the principles in the psychology of sensation that are satisfied by the genuine sensing relation. Therefore, these psychological principles do not implicitly define the sensing relation and, hence, do not form the basis for a direct functional definition of sensing. However, this line of criticism is inconclusive. First, the possibility that sense experience is "uninterpreted" or "raw" might mean that sensing is not an intentional relation. If so, the inverted-spectrum example would leave untouched the functional definability of intentional relations, which is perhaps the most interesting aspect of functionalism. Second, there are principles of traditional philosophical psychology and epistemology that sensing* cannot satisfy, namely, those concerning introspective intentional states and states involving direct experience of the conscious operations of mind themselves. (See pp. 300-304 for a fuller discussion of such principles.) According to one such principle, for example, although a person can introspect that he is sensing red, no one could introspect that he is sensing* red. For whereas sensing is a conscious operation of mind, the *ad hoc* relation sensing* is not, and thus, unlike the genuine conscious operations of mind, it is not open to direct introspective scrutiny. In this connection, a person's knowledge that he is sensing red need not be inferential in any sense, but his knowledge that he is sensing* red is always inferential in some sense. Or consider another example. According to traditional empiricist psychology (such as that of John Locke), a person can in acts of reflection actually experience his own conscious operations of mind (sensing, feeling, thinking, craving, introspecting, and experiencing

itself), but he cannot in any way experience such *ad hoc* operations as sensing*. The point here is not that principles such as these are right; the point is that they might be right. So unless the critic can *show* once and for all that they are not, he has not refuted functionalism. Since this is out of the question, the only way for the critic to block a response based on principles such as these is to extend the "inversion" strategy *globally* to all other mental relations. That is, the critic must construct, not only the sensing* relation, but also the relations feeling*, thinking*, craving*, introspecting*, experiencing*, etc., which are synchronized with sensing*, and then he must show that, taken together, these relations satisfy the full body of psychological principles contemplated by the sophisticated functionalist. For example, he must show that a person can introspect* that he is sensing* red and that a person can experience* such things as sensing*, feeling*, thinking*, introspecting*, and even experiencing*. Until now no critics have attempted this kind of global counterexample, but only this will suffice.[13]

The final type of counterexample I will consider is more or less parallel to the inverted-spectrum example except that it is aimed at intentional relations rather than the sensing relation. Searle's Chinese-room example falls into this type of criticism.[14] According to Searle, someone who is locked in a room and who knows no Chinese could, by following appropriate instructions written in English, instantiate a computer program for "speaking Chinese" and yet fail to be speaking Chinese. Searle concludes that a certain narrow version of functionalism, namely, Turing-machine functionalism, must be mistaken. (Like more liberal versions of functionalism, Turing-machine functionalism–or AI functionalism–holds that the standard mental relations are implicitly defined by the principles of psychology. But it goes further, requiring that the principles of psychology are, in effect, no more than a Turing-machine table, i.e., an abstract characterization of a purely mechanical computational process.) Perhaps, when suitably refined, the Chinese-room example can refute certain forms of Turing-machine functionalism, but can it refute functionalism per se? After all, there are many forms of Turing-machine functionalism, and there are many versions of functionalism besides Turing-machine functionalism.[15]

To be a counterexample to functionalism per se, the activity of the person in the Chinese room would have to be functionally indistinguishable from the activity of speaking Chinese. But a functionalist has excellent grounds for denying that it is. For example, the functionalist can hold, à la Grice and Searle, that speaking Chinese is a complex intentional activity, namely, an activity performed with the intention to impart beliefs to hearers who are intended to recognize that original intention via a certain inference route involving a certain mutually agreed upon system of rules (i.e., the semantics and pragmatics for Chinese). Now in order for the Chinese-room activity to be functionally indistinguishable from this complex intentional activity, it is not enough that the phonetic or orthographic input/output functions should match. The inner states must be functionally isomorphic as well. How might we fill out the Chinese-room example to meet this essential further condition? There are really only two alter-

natives. First, we could simply specify that the Chinese-room activity be performed with all the standard Gricean intentions. However, in that case the activity would be just one more instance of speaking Chinese (though admittedly a recherché one), so it would be no counterexample. The second alternative would be to try to construct new relations (intending*, believing*, recognizing*, inferring*, mutually agreeing*) that are distinct from, but functionally isomorphic to, the standard intentional relations (intending, believing, recognizing, etc). And then we might specify that the Chinese-room activity be performed with the intention* to impart beliefs* to hearers who are intended* to recognize* that original intention* via a certain inference* route involving a certain mutually agreed* upon system of rules (i.e., the semantics and pragmatics for Chinese). In this case, the requisite isomorphism would be insured, and yet the Chinese-room activity would not qualify as speaking Chinese. But filling out the counterexample in this way would require establishing the existence of the new relations intending*, believing*, recognizing*, etc., which are functionally isomorphic to the standard relations of intending, believing, recognizing, etc. Notice, however, that these standard relations participate in a wide variety of psychological principles, principles that involve in one way or another *every* basic mental relation, including even introspecting and experiencing (discussed above in connection with sensing). Moreover, these psychological principles are thought by many (especially by traditional philosophical psychologists) to have a strong modal value. (For example, some think that these principles hold, not just in all actual situations, but in all causally possible counterfactual situations as well. Others think that the basic psychological principles are constitutive of rationality and, therefore, that, for all normal beings in all normal circumstances, these principles are metaphysically necessary; and others inclined to strict essentialism about the mind think that the basic principles are metaphysically necessary with no qualification vis-à-vis normality.) Therefore, to fill out the Chinese-room example, we must establish the existence of a *global* system of relations–intending*, believing*, recognizing*, sensing*, feeling*, craving*, introspecting*, experiencing*, etc.–that satisfies the *whole* body of psychological principles, perhaps with a strong modal value. If we could do this, though, we would thereby already have succeeded in refuting functionalism, and the Chinese room would drop out of picture. Furthermore, if true psychological principles really do have strong modal values, it is just not plausible that artificial, contingent set-ups such as the Chinese room or ones patterned after it could yield a global system of deviant relations that could satisfy such modally qualified principles. Moreover, this conclusion would hold even if these principles stipulate that all mental processes must be computational in nature.

There are clear flaws, then, in all the current candidate counterexamples. They all fall short of a refutation of functionalism because they fail to provide a global system of relations that uncontroversially can be shown to differ from the standard mental relations and that satisfy the whole body of psychological principles, principles that might well possess strong modal values. So how are we to go about constructing the

requisite global system of deviant relations, intending*, thinking*, etc.? The example that is heuristically useful for this purpose is the inverted spectrum. Recall that the function f that inverts the color spectrum by mapping it onto itself can be used to define a nonstandard relation sensing* which functions (at least locally) rather like the standard sensing relation. On analogy, suppose we can find a way to "invert" our total conceptual scheme, i.e., a function that maps our total conceptual scheme onto itself. Then, this function can be used to define a system of "inverted" relations–intending*, thinking*, etc.–that correspond to the system of standard mental relations–intending, desiring, etc. If these definitions are suitably constructed, the system of inverted relations will satisfy the body of psychological principles that, according to functionalism, implicitly defines the standard mental relations. However, if there are two systems of relations–the standard mental relations and their inverted counterparts–that satisfy this body of principles, then these principles cannot implicitly define the standard mental relations, and, hence, they cannot provide the basis for direct functional definitions of the standard mental relations. Thus, the existence of these two systems of relations–that is, the existence of mind and anti-mind–will lead to a formal disproof of the central thesis of functionalism; the principles of psychology will be blind to the distinction between mind and anti-mind.

THE ELEMENTARY BELIEF/DESIRE MODEL

When one thinks of inversions of the conceptual scheme, Quine's thesis of the indeterminacy of translation springs to mind. Quine's indeterminacy thesis derives ultimately from the insight that speakers of a radically alien language could think–or could be interpreted as thinking–in a way that is systematically different from ours and yet that is fully as rational. And this insight inspires Quine's premise that there exists more than one fully adequate translation manual for each radically foreign language. Now Quine uses this premise to support his conclusion that there is no objective fact of the matter concerning which translation manual is correct, and, by parity of reasoning, Quine would use the original insight to support the conclusion that there is no objective fact of the matter concerning which way the foreigners really think, our way or the systematically different way. But these two conclusions need not concern us now (though I will return to them later). What matters at present are the Quinean premises that speakers of a radically alien language could think–or could be interpreted as thinking–in a way systematically different from, but fully as rational as, the way we think and, relatedly, that there exists more than one fully adequate translation manual for their language.

Let L be a radically alien language for which there are two fully adequate translation manuals, and let g_1 and g_2 be the functions, generated by these manuals, that map sentences of L onto their English "translations." For example, perhaps g_1('Gavagai gua') = 'There exists a rabbit' and g_2('Gavagai gua') = 'Rabbithood is manifest'. (N.B. For ease of presentation I will use this particular/universal example

in my preliminary remarks about Quinean transformations. Later in this section I will also use a certain left/right transformation, which was inspired by conversation with Donald Davidson and with George Myro.) Let the function g, which maps English sentences onto English sentences, be defined as follows: $g(A) =_{df} g_2(g_1^{-1}(A))$. That is, the value of g when applied to the English sentence A is the English sentence that results when g_2 is applied to the L-sentence whose g_1-value is A. For example, g('There exists a rabbit') = 'Rabbithood is manifest'. And if g_1 and g_2 have been suitably constructed, g('Rabbithood is manifest') = 'There exists a rabbit'. Indeed, if g_1 and g_2 have been suitably constructed, the following will hold for all English sentences A: $g(g(A)) = A$. Now let m be the function that maps English sentences onto their meanings, i.e., onto the propositions they express. For example, since 'There exists a rabbit' means that there exists a rabbit, m('There exists a rabbit') = the proposition that there exists a rabbit. Then we define a function t that maps propositions onto propositions: $t(p) =_{df} m(g(m^{-1}(p)))$. That is, the value of t when applied to proposition p is the proposition that is expressed by the sentence that is the g-value of the sentence that expresses p. For example, t (the proposition that there exists a rabbit) = the proposition that rabbithood is manifest. And if g_1 and g_2 have been suitably constructed, t(the proposition that rabbithood is manifest) = the proposition that there exists a rabbit. Indeed, for any proposition p expressible in both English and L, $t(t(p)) = p$. That is, t maps each proposition expressible in both English and L to an inverted counterpart, and each inverted counterpart, right back to the original.

The function t has been constructed in a linguistic setting, and accordingly t is defined only on propositions expressible in both English and the alien language. This restriction can easily be removed, however, by skipping the linguistic detour and constructing t directly within a theory of properties, relations, and propositions. In this construction t would be defined on everything in the domain of discourse–individuals, properties, relations, and propositions. Further, $t(t(x)) = x$ would hold for all x in the domain. This function t would thus invert the conceptual scheme in the way we are seeking. In what follows I will call functions of this kind *Quinean transformations*. (Incidentally, our construction would go through substantially unchanged if, as David Lewis suggests in "General Semantics," we were to identify the objects of the propositional attitudes with a certain kind of abstract tree rather than with propositions.)

We can use a Quinean transformation t to define an inverted relation for each of the standard propositional attitudes:

x thinks* p iff$_{df}$ x thinks $t(p)$.[16]

x wants* p iff$_{df}$ x wants $t(p)$.

. . . .

With these inverted relations in hand, we can now return to functionalism. In this section I will assess the adequacy of functionalism relative to a certain set of psychological principles, viz., the elementary belief/desire model of action (described below). (I do the same for a more sophisticated set of principles in the next section.) My assessment follows from two theses. First, the system of inverted relations (thinks*, etc.) resulting from the above definitions satisfies all the principles in the elementary belief/desire model of action. Second, (at least some of) these inverted relations are significantly different from their standard counterparts; indeed, they are typically not even co-extensive with them. Given these two theses, it follows that the elementary belief/desire model of action does not implicitly define (all of) the standard propositional attitudes and, therefore, that this model does not provide the basis for direct functional definitions of them either. This conclusion is of interest for two reasons. First, many cognitive psychologists and philosophical functionalists are content with this model. And second, though the model might not reflect all of the distinctive features of a sophisticated mental life, if "lower" creatures (and perhaps also computers) have minds, their minds might be adequately described by a model with this degree of complexity.

I begin with a defense of the first thesis. My strategy shall be to present an idealization of the elementary belief/desire model of action and then to prove that, if the standard propositional attitudes satisfy this idealized model, then so do their inverted counterparts. Then I will indicate why this result can be extended to richer versions of the elementary belief/desire model.

When I speak of the elementary belief/desire model of action, I mean the psychological theory consisting of all elementary principles for the standard propositional attitudes. A principle for the attitudes is elementary if it is not concerned with the behavior of the attitudes in connection with multiple embeddings. Thus, principles such as the following are *not* elementary: (Infallibility) If x is thinking that x is thinking p, then x is thinking p. I will deal with psychological models that include non-elementary principles in the next section.

The following will be our idealization of the elementary belief/desire model. I will take the liberty to simplify significantly the principles contained in it, but nothing substantive should ride on this. For all x, p, q, and F:

Input. If p is true and it is causally necessary at the moment that x entertains p if and only if p is true, then x thinks p.[17]

Pure Reason. If x thinks that if p then q and x thinks p and x entertains q, then x thinks q.

Practical Reason. If x thinks that if p then q and x wants q and x does not want not p and x entertains p, then x wants p.

Intention. If x wants that Fx and x thinks that x can F at the moment, then x intends that Fx.

Output. If x intends that Fx and it is causally possible at the moment that Fx, then Fx.

The thesis that the system of inverted relations–thinking*, wanting*, etc.–satisfies these principles naturally depends on our choice of Quinean transformation t. I will suggest two transformations that easily fill the bill (though there are many

others). The first, which I used in my illustrations earlier, is a function that maps existential propositions (e.g., that there exists a rabbit) to propositions concerning the manifestation of associated universals (e.g., that rabbithood is manifest) and propositions concerning the manifestation of universals back to associated existential propositions. I call this the *particular/universal transformation*. Since it is fairly obvious how to define this transformation rigorously, I will suppress technicalities until the next section when they are unavoidable. Suffice it to say that, aside from converting existential propositions into propositions concerning the manifestation of associated universals (and conversely), the particular/universal transformation leaves everything else about a proposition unchanged. That is, this transformation leaves intact all the "nonformal" features of a proposition. For example, it does not alter any proposition that is embedded as a subject within a given proposition upon which it is acting.

The second transformation, which I call the *left/right transformation*, is constructed as follows. (The construction invokes a certain kind of artificial object that is a logical construct out of ordinary physical objects. Such artificial objects could be avoided by complicating the construction in certain ways.)[18] Let $l(x)$ be the infinite moving plane that always follows x around dividing his body (and the universe) into a left half and a right half. Let the function h_x be defined as follows. If $v = x$ or v is not located in space, then $h_x(v) =_{df} v$. Otherwise, $h_x(v) =_{df}$ the object u such that, necessarily, if x exists, u is located at the same distance d ($d \geq 0$) from the plane $l(x)$ as v but on the opposite side, and if x does not exist, u has the same location as v. Thus, if x exists, $v \neq x$, and v is located in space, then $h_x(v)$ is v's "logical shadow" on the other side of the plane $l(x)$.[19] Let us extend h_x to properties and i-ary relations as follows: $h_x(F)$ = the property of being a v such that $F(h_x(v))$ and $h_x(R)$ = the relation holding among $v_1, \ldots, v_i$ such that $R(h_x(v_1), \ldots, h_x(v_i))$. Then the transformation t_x may be defined inductively. To illustrate, I will state the definition for the case of atomic propositions: if $h_x(F) = F'$, $h_x(R) = R'$, and $h_x(a_k) = a_k'$, $k \geq 1$, then $t_x(F(a_1)) = F'(a_1')$ and $t_x(R(a_1, \ldots, a_i)) = R'(a_1', \ldots, a_i')$. More complex cases are handled analogously. (Naturally, our definitions of the inverted relations are to be understood as having the following form: x thinks* p iff$_{df}$ x thinks $t_x(p)$. That is, a person thinks* p if and only if he thinks the proposition that arises from p via the left/right transformation that is defined in terms of the infinite plane that always moves around with his own body dividing it left and right. Incidentally, when the context permits, I will often write t instead of t_x in order to simplify things notationally.) To get a better idea of how the left/right transformation works, consider the proposition that this (i.e., my left hand) moves. Let h_{me}(this) = this' [i.e., my left hand's "logical shadow" on the other side of the plane l(me)]. Let h_{me}(moving) = moving' [roughly, the property of being the "logical shadow" of something that moves on the other side of the plane l(me)]. Then, t_{me}(the proposition that this moves) = the proposition that this' moves'.

The particular/universal transformation and the left/right transformation are defined so that this lemma follows directly: for any proposition p, p and $t(p)$ are necessarily equivalent. Given this lemma, it is straightforward to prove that the system of

inverted relations based on either Quinean transformation satisfies the principles in our idealization of the elementary belief/desire model. For example, consider the Input principle. We wish to show that the inverted relations thinking* and entertaining* satisfy this principle. It will suffice to show that the following inverted counterpart of the original principle holds for arbitrary p:

> If p is true and it is causally necessary at the moment that x entertains* p if and only if p is true, then x thinks* p.

Suppose that $t(p) = q$. Then, given the definitions of 'thinks*' and 'entertains*', this principle is necessarily equivalent to the following:

> If p is true and it is causally necessary at the moment that x entertains q if and only if p is true, then x thinks q.

Notice that 'p' now occurs only in contexts in which necessary equivalents can be validly substituted *salva veritate*. However, by our lemma, we know that p and q are necessarily equivalent. So 'q' may be validly substituted for 'p' *salva veritate*. Therefore, the last formula–and, in turn, the inverted counterpart of the original Input principle–is necessarily equivalent to the following:

> If q is true and it is causally necessary at the moment that x entertains q if and only if q is true, then x thinks q.

But this is just an instance of the original Input principle. Therefore, since the original Input principle holds, it follows that the inverted counterpart of this principle holds too. In the case of the Intention and Output principles the proofs are much the same except that, when dealing with the particular/universal transformation, they make use of the facts that t(the proposition that Fx) = the proposition that Fx and t(the proposition that x can F at the moment) = the proposition that x can F at the moment. And when dealing with the left/right transformation, they make use of the facts that t_x(the proposition that Fx) = the proposition that $F'x$ and t_x(the proposition that x can F at the moment) = the proposition that x can F' at the moment. Finally, in the case of the Pure Reason and Practical Reason principles, the proofs are again much the same except that they make use of the fact that t(the proposition that if p then q) = the proposition that if r then s, where $r = t(p)$ and $s = t(q)$.

So, therefore, the systems of inverted relations based on our two Quinean transformations satisfy our idealization of the elementary belief/desire model of action. Moreover, this conclusion generalizes. For example, let T be any elementary principle for the standard propositional attitudes such that (1) every formula within the scope of a propositional-attitude predicate in T is built up out of propositional variables and propositional connectives, and (2) except for propositional-attitude predicates, each constant in T is either a predicate of individuals or a logical operator for which the principle of the substitutivity of necessary equivalents is valid.[20] (These operators may include, e.g., modal, causal, probability, and even counterfactual operators as long as the substitutivity principle is valid for them.) Let T^* arise from T by replacing all

predicates for the standard propositional attitudes with predicates for their inverted counterparts. Then our generalization is this: T and T^* are necessarily equivalent, and, hence, if P is any infinite class of T-like principles and P^* the associated class of T^*-like principles, P and P^* are necessarily equivalent. This generalization is proved by a straightforward inductive argument.

Stronger generalizations are also possible. In proofs of these further generalizations the following resources are used. First, if t is the particular/universal transformation, all the nonformal "constituents" in $t(p)$ are the same as those in p, and since p and $t(p)$ arise from one another by a mechanical logical manipulation, they are not just necessarily equivalent but are provably equivalent (in a theory of properties, relations, and propositions). Second, if t is the left/ right transformation, p and $t(p)$ always have the same logical form. Third, consider the Quinean premise discussed at the outset of this section, i.e., the premise that there could be beings who think (or can be interpreted as thinking) in systematically different, yet equally rational, ways. When this premise is rigorously formulated, it entails that, for every principle characterizing some feature of one way of thinking (e.g., "particularese"), there exists a complementary principle characterizing an analogous feature of a corresponding way of thinking (e.g., "universalese"). Take as an example the following "particularese" principle:

> If x clearly and distinctly understands the proposition that (Fa only if something is F), then x will think that (Fa only if something is F).

The complementary "universalese" principle is this:

> If x clearly and distinctly understands the proposition that (Fa only if F-ness is manifest), then x will think that (Fa only if F-ness is manifest).

Clearly, if the original principle is true, so is the complementary principle. To show that the inverted relations satisfy the first principle, replace 'understands' with 'understands*' and 'think' with 'think*'. Then, expand 'understands*' and 'think*' in accordance with their definitions. After that, replace 't(that Fa only if something is F)' with the necessarily equivalent expression 'that Fa only if F-ness is manifest'. The result is none other than the complementary principle. Since the latter principle is true and since our substitutions are equivalence preserving, it follows that the inverted relations satisfy the original principle. Moreover, by analogous argument, we can employ the original principle to show that the inverted relations also satisfy the complementary principle. More generally, if the standard psychological attitudes satisfy any universal psychological principle–i.e., any psychological principle that is neither species-dependent nor particular-dependent (see below)–then this fact can be used to show that the inverted relations satisfy the complementary principle, and conversely.[21] This special complementarity, together with the earlier two facts, entail much stronger generalizations on our previous results. Indeed, these generalizations extend to all universally applicable elementary psychological principles, i.e., to all

elementary principles that concern the *general* nature of the mind rather than only the mental idiosyncrasies true of isolated species or of isolated individuals.

Granted that the inverted relations satisfy all true universal elementary principles, there are, nevertheless, true species-dependent and particular-dependent principles that the inverted relations do not satisfy.[22] However, this fact can do nothing at all to save functionalism. First, given that there is an open-ended list of possible species and possible thinking beings, species-dependent and particular-dependent principles have no place in the philosophical analysis of a universally applicable concept such as thinking. In this connection, one of the fundamental goals of functionalism is to provide a *general explanation* of why psychological concepts are true of some entities and not of others. If species-dependent or particular-dependent principles are invoked in the functional definitions of psychological concepts, this fundamental goal of functionalism is completely undermined. Second, consider any set of species-dependent (or particular-dependent) principles there might be for any given list of possible species (or particulars). Even if such principles are used as the basis for a functional definition of, e.g., thinking, this functional definition will always fail, for such principles will always be satisfied by some new "inverted" relation (e.g., thinking**) which we can construct by adapting the formal technique sketched at the close of this section.

So our overall conclusion is this. Any system of universal (i.e., non-species-dependent and non-particular-dependent) elementary principles is satisfied by the standard psychological attitudes only if it is also satisfied by one of our systems of inverted relations; moreover, any system of elementary principles containing possibly the non-universal principles for any list of possible species (or particulars) is satisfied by the standard psychological attitudes only if it is also satisfied by some new system of inverted relations constructed in accordance with the technique mentioned just above.

A functionalist who is content to work with the elementary belief/desire model of action might hope to avoid this conclusion by emphasizing the notion of causal role.[23] Even if the inverted relations and the standard relations satisfy the same causal laws, perhaps they do not have the same causal roles. However, relations, just on their own, have no causal efficacy at all. Events cause events; phenomena cause phenomena. Relations cause neither. Now it is true by *definition* that xS^*p if and only if $xSt(p)$, for any standard propositional attitude S and its inverted counterpart S^*. For example, it is true by *definition* that you are thinking* p if and only if you are thinking $t(p)$. Therefore, the event of someone's thinking* p must be identical to the event of that person's thinking $t(p)$. In general, every event of thinking* must be identical to an event of thinking. Furthermore, it is *provable* that, for every proposition q, there is a p such that, necessarily, $q = t(p)$ and, hence, necessarily, x thinks q if and only if x thinks $t(p)$.[24] Therefore, even on extremely strict criteria of event identity (criteria much stronger than those used in arguments for the token/token identity thesis), for every q, there is a p such that the event of someone's thinking q is identical to the event of that person's thinking $t(p)$. However, we have just seen that the event of

someone's thinking $t(p)$ is identical to the event of that person's thinking* p. It follows, therefore, that for every proposition q, there is a p such that the event of someone's thinking q is identical to the event of that person's thinking* p. Thus, every event of thinking is identical to an event of thinking*. So, in summary, every event of thinking* is an event of thinking, and every event of thinking is an event of thinking*. It follows that any event of thinking* possesses exactly the same causal role as an event of thinking and any event of thinking possesses exactly the same causal role as an event of thinking*. It is implausible, therefore, that the usual notion of causal role could help save functionalism from the line of criticism developed in this section.

Let us turn finally to my second thesis, namely, that (at least some of) the inverted relations thinking*, wanting*, etc. are significantly different from their standard counterparts thinking, wanting, etc. There is compelling evidence for this thesis. Consider thinking, for example. It is clearly possible for a person to be thinking that there exists a rabbit and to fail to be thinking that rabbithood is manifest. For that matter, most people who are right now thinking that there exists a rabbit are not also thinking that rabbithood is manifest. Indeed, many people who have thought the former have never thought the latter, and some are not disposed even to *understand* the latter.[25] However, the latter proposition is the result of applying the particular/universal transformation to the former proposition. Therefore, there is a proposition p such that someone thinks p but does not think $t(p)$, i.e., such that someone thinks p but does not think* p. So, when thinking* is defined in terms of the particular/universal transformation, thinking and thinking* are not even materially equivalent. (For further justification of this conclusion, see below.)

Next consider the case where thinking* is defined in terms of the left/right transformation. Take the proposition that this (i.e., my left hand) moves. Let h_{me}(this) = this′ [i.e., my left hand's "logical shadow" on the other side of the plane l(me)]. Let h_{me}(moving) = moving′. Then t_{me}(the proposition that this moves) = the proposition that this′ moves′. Plainly, it is possible for me to think that this moves without thinking that this′ moves′. Moreover, I never thought that this′ moves′ until today even though I have numerous times thought that this moves. Indeed, I confess to having some difficulty understanding with clarity and distinctness the proposition that this′ moves′. Thus, there is plainly a proposition p such that I have thought p and not thought $t_{me}(p)$. Therefore, when thinking* is defined in terms of the left/right transformation, thinking and thinking* again are not even materially equivalent. So, with either of our two Quinean transformations, thinking and thinking* are significantly different.

Are there any grounds on which a functionalist might resist this conclusion? There are three that I can think of, and I will discuss them in turn. First, there is a theory according to which all intensional entities, including propositions, are identical if necessarily equivalent. (The familiar possible-worlds treatment is committed to this theory.) However, our Quinean transformations have been constructed so that, for any proposition p, $t(p)$ and p are necessarily equivalent. Therefore, according to the above theory of the identity conditions for propositions, $t(p) = p$. Hence, it would follow that, if we use these transformations in our definitions, thinking = thinking*, desiring = desiring*, etc. And this would defeat my argument.

However, most people today acknowledge the striking inadequacy of this theory of the identity conditions for propositions. For example, if the theory is right, then whoever knows a trivial necessary truth knows every necessary truth, and this is plainly false. Or consider an argument involving ontological commitment. Take the proposition that there exists a rabbit and the proposition that rabbithood is manifest. If the theory is right, these two propositions would have the same ontological commitments. But clearly they do not. The former proposition is not committed to the existence of any universals, but the latter proposition is. This point can be dramatized as follows. An extensionalist philosopher who, like Quine, believes that universals do not exist would deny the latter proposition and yet affirm the former. If the above theory were correct, such a philosopher would be guilty of explicit contradiction. But obviously this is not the case. Finally, the best functional explanations of certain kinds of verbal behavior invoke fine-grained intensional distinctions. Suppose that I am a normal, sincere English speaker. Then the most straightforward functional explanation of my assertion of the sentence 'Awhile ago I was thinking that there exists a rabbit but was not thinking that rabbithood is manifest' requires that there be a distinction between thinking that there exists a rabbit and thinking that rabbithood is manifest. Do those who are realists about the propositional attitudes have any good reason not to accept such an explanation?

For these sorts of reasons, we may safely conclude that the standard propositional attitudes characteristically are sensitive to distinctions cut more finely than necessary equivalence. Those who do not honor such distinctions might be interested in relations having certain gross resemblance to the standard propositional attitudes, but they certainly are not interested in these relations themselves.

The second reason someone might object to our conclusion that thinking ≠ thinking* goes as follows. Quine uses his indeterminacy argument to urge that there exists no objective difference between an alien's meaning p and his meaning $t(p)$ and that there exists no objective difference between an alien's thinking p and his thinking $t(p)$. Therefore, it is improper for us to use a Quinean transformation to define the relation thinking*–thinking* $p =_{df}$ thinking $t(p)$–and then to go on to claim that there is an objective difference between thinking and thinking*. One way to reply to this objection is to challenge Quine's conclusion. That is, we could accept Quine's premise that it is impossible for us to determine whether an alien means p rather than $t(p)$ and that it is impossible for us to determine whether an alien is thinking p rather than $t(p)$; and at the same time, we could deny that it follows from this that there is no objective difference between an alien's meaning p and his meaning $t(p)$ or between his thinking p and his thinking $t(p)$. However, each of us on his own can avoid this head-on challenge to Quine's conclusion simply by shifting to the *first person*.[26] For example, *I* know that thinking ≠ thinking* if any of the following holds. At least once

I have been thinking that there is a rabbit and have failed at that time also to have been thinking that rabbithood is manifest. Or at least once I have been thinking that something hits something and have failed at that time also to have been thinking that the property of being something x such that the property of being hit by x is manifest is itself a manifest property. (See note 25.) Or at least once I have been thinking that this (i.e., my left hand) moves and have failed at that time also to have been thinking that this′ moves′. And I know I have. So I know that thinking $\neq$ thinking*. Therefore, I may confidently conclude that no elementary psychological principles can implicitly define what it is for me to be thinking p rather than to be thinking* p. Likewise, each of you can go through an analogous chain of reasoning, at the end of which you can conclude that no elementary psychological principles can implicitly define what it is for you to be thinking p rather than to be thinking* p.

The third possible objection to the conclusion that thinking $\neq$ thinking* is that the construction of thinking* presupposes that thinking is a relation holding between individuals and propositions; functionalists who believe that the thinking relation holds between individuals and pure syntactic entities such as sentences in Mentalese (''the language of thought'') might make this sort of objection. However, this whole issue can be side-stepped by re-defining our Quinean transformation t so that it maps, e.g., sentences in Mentalese onto corresponding necessarily equivalent, non-synonymous sentences in Mentalese. (Our function g from the beginning of this section is just like this re-defined function t except that it maps English sentences onto necessarily equivalent, non-synonymous English sentences.) Once t is re-defined in this way, the rest of the above argument goes through substantially unchanged.

The ''syntactic functionalist'' might believe, however, that this re-definition of t gives him a special advantage not available to the ''propositional functionalist.'' The reason is that, since syntactic entities are pure abstract shapes, they can have physical inscriptions in such materials as brain cells; propositions cannot. (To assist the syntactic functionalist, I will use the term 'pure abstract shape' in a liberal way so that quite dissimilar physical objects may, if desired, be counted as having the same pure abstract shape. However, the use of this term should not be so liberal that a sentence and its Quinean transform are counted as being the same pure abstract shape, for that would undermine the syntactic functionalist's ability to preserve the sort of fine-grained distinctions that are characteristic of intentionality.) So, for example, since a given Mentalese sentence S differs in pure abstract shape from its transform $t(S)$, the inscriptions of S and of $t(S)$ in the brain will differ physically. Now, the syntactic functionalist might think that such physical differences will in turn help to distinguish the standard mental relations from their inverted counterparts. However, as I will show, this is not so.

Suppose for the sake of argument that the objects of the standard psychological attitudes are not propositions but sentences in Mentalese and that t has been re-defined so that it maps Mentalese sentences onto appropriate necessarily equivalent, non-synonymous Mentalese sentences. Now functionalism is based on the insight that functionally indistinguishable structures can have quite different physical realizations. The only thing that two such realizations must have in common is the following: (1) there must be a structural isomorphism between the components of one and the components of the other, and (2) the components of one must interact causally with one another and their outside environment in the same way as the components of the other interact with one another and their outside environment. Beyond this it makes no difference what the components are like. In particular, it makes no difference just which syntactic entities (i.e., which pure abstract shapes) are physically inscribed in which realization. All that is required is that each syntactic entity physically inscribed in one realization have a structural isomorphism to the corresponding syntactic entity physically inscribed in the other realization and that the causal role played by inscriptions of the former within the first realization is the same as the causal role played by inscriptions of the latter within the second realization. However, physical inscriptions of sentences that are Quinean transforms of one another could easily have such a structural and causal isomorphism to one another. After all, sentences that arise from one another via the left/right transformation are identical in syntactic form, and sentences that arise from one another via the particular/universal transformation differ only by a syntactic manipulation that, from a mechanical point of view, has no more significance than a difference in punctuation convention.

To dramatize the point, consider the following rather fanciful hypothetical example.[27] According to functionalism, it should be possible that each being of type-a thinks the Mentalese sentence S if and only if an actual inscription of, say, the English sentence A occurs in a certain region of its brain and that it thinks $t(S)$ if and only if an inscription of $g(A)$ occurs in that region. And according to functionalism, since A and $g(A)$ have a direct structural isomorphism to one another (i.e., the function g itself), it should also be possible that each being of type-b (perhaps a type of being made of quite different materials) thinks S if and only if an inscription of $g(A)$ occurs in an analogous region of its brain and that it thinks $t(S)$ if and only if an inscription of A occurs in that region of its brain. Now according to functionalism, since A and $g(A)$ have a direct structural isomorphism to one another (i.e., function g itself), the causal role of inscriptions of A within type-a beings would be identical to the causal role of inscriptions of $g(A)$ within type-b beings, and the causal role of inscriptions of $g(A)$ within type-a beings would be identical to the causal role of inscriptions of A within type-b beings. Thus, according to functionalism, beings of type-a would be functionally indistinguishable from beings of type-b. Therefore, the mere existence of A-inscriptions or of $g(A)$-inscriptions in a being's ''thinking center'' does nothing in itself to indicate whether a being is thinking S or whether it is thinking $t(S)$ instead; i.e., it does not indicate whether the being is thinking S or whether it is thinking* S. Generalizing on this, we may conclude that the fact that syntactic entities can have physical instances in brains does nothing to reduce the threat that thinking* satisfies all the principles of functional psychology that are satisfied by thinking.

The functionalist might hope to avoid this conclusion by retreating from pure

functionalism and by moving back toward naive physiological reductionism. To do this, he would add to his set of psychological principles a list of species-dependent psycho-physiological correlations, for example:

the type-a-being-thinks-S/A-inscription correlation;
the type-a-being-thinks-$t(S)/g(A)$-inscription correlation;
the type-b-being-thinks-$S/g(A)$-inscription correlation;
the type-b-being-thinks-$t(S)/A$-inscription correlation.

But even this will not help. We can show that the enriched set of principles does not implicitly define thinking, desiring, etc. by defining a new system of inverted relations thinking**, desiring**, etc. that are tailored to the enriched set. For example, x thinks** S iff$_{df}$ x thinks S if x is type-a or type-b and otherwise x thinks* S. So defined, thinking** and thinking, desiring** and desiring, etc. are significantly different; indeed, they can fail to be co-extensive. For according to functionalism, there is an *open-ended* list of different possible physical realizations of the same functional structure; type-a and type-b beings are only two of a potentially infinite number of such realizations. Although the extension of thinking** and the extension of thinking are the same concerning beings of type-a and type-b, concerning all these other types of beings they will differ. But do thinking**, desiring**, etc., satisfy the enriched set of principles? Since thinking*, desiring*, etc., satisfy the original set of principles (and we have already shown that they do), thinking**, desiring**, etc. satisfy the enriched set. For thinking and thinking** must have the same extensions over type-a and type-b beings, and thinking* and thinking** must have the same extensions over all other types of beings.[28]

In this way, we can always keep one step ahead of the functionalist who would add psycho-physiological correlations to his other psychological principles: for any enriched set of principles satisfied by the standard propositional attitudes, we can define a system of inverted relations that also satisfies them. Therefore, invoking the "language of thought" or other syntactical constructs does nothing to ward off the line of attack we have mounted against functionalism. Thus, as far as the truth of functionalism is concerned, we might as well identify propositions as the objects of the standard propositional attitudes; the syntactical alternatives offer no advantage.[29]

This completes my discussion of possible objections to my argument. The overall conclusion is that the elementary belief/desire model of action cannot implicitly define the standard propositional attitudes and, hence, cannot serve as the basis of direct functional definitions of them either. However, we have not shown that stronger psychological principles cannot validate functionalism. To show this, we must turn to the principles that characterize the self-conscious rational mind.

THE SELF-CONSCIOUS RATIONAL MIND

The elementary belief/desire model contains no principles concerned with the behavior of the propositional attitudes in connection with multiple embeddings and no

principles concerned with the behavior of psychological relations that are not propositional attitudes. When such principles are added to the elementary belief/desire model, we approximate a theory adequate for characterizing the self-conscious rational mind. Perhaps this enriched theory–or a perfected version of it–is sufficiently strong to define implicitly the standard psychological relations and, in turn, to provide a basis for direct functional definitions of them. Our previous construction, at any rate, does not rule out this possibility, for the system of inverted relations (thinking*, etc.) turns out not to satisfy many of the principles that might be added when we enrich the belief/desire model. What we shall discover is this. Intensionality–the very phenomenon that made our previous "Quinean" construction succeed in the elementary setting–leads to its downfall in the more sophisticated setting of the self-conscious rational mind. Specifically, there are intensional distinctions that show up in the way the mind gains access to information about itself–to the *contents* of its own conscious intentional states and to its own conscious *operations*–and our "Quinean" construction is insensitive to these distinctions. Thus, the prospect of a successful functionalism is still alive.

Consider the matter of the contents of one's own conscious intentional states. Each of our Quinean transformations–and each Quinean transformation known in the philosophical literature–is "superficial" in the sense that it transforms only the "uppermost" level of any proposition upon which it is operating; it leaves intact the propositions embedded as subjects within those propositions. (For example, the particular/universal transformation simply alters the uppermost quantificational structure of a proposition p; it does not alter the quantificational structure of propositions embedded within p.) However, non-elementary principles that characterize the mind's access to the contents of its own intentional states typically "mix levels." That is, they describe direct internal links between, e.g., x's thinking a proposition p and x's thinking the proposition that x is thinking p. (It is crucial that p be the same in each case.) When a Quinean transformation t is applied to a proposition p, often the result q is significantly different. However, when t is applied to the proposition that x is thinking p, the embedded proposition p is left intact; the new proposition q does not take its place. (For example, the result of applying the particular/universal transformation to the proposition that x is thinking p just is the proposition that x is thinking p. This is so because this proposition is "atomic" and has no quantificational structure. The fact that p might itself have a complex quantificational structure is beside the point.) Now although p and q are necessarily equivalent, there is always a significant fine-grained intensional information gap between propositions in which p is embedded and those in which q is embedded. Therefore, the familiar "superficial" Quinean transformations break the direct internal link between x's thinking p and x's thinking that x is thinking p. And since our inverted relations thinking*, etc. are defined in terms of such Quinean transformations, they too fail to preserve the direct internal links that typify the self-conscious rational mind. Indeed, to restore this kind of direct internal link, we will be forced to construct an entirely new system of inverted relations.

I will illustrate the breakdown in our earlier construction by examining in more detail some principles that have played an important role in the history of philosophical psychology and epistemology. For convenience, I will formulate these principles in a simplified–and, no doubt, overly strong–way, and I will employ the traditional idiom.

The first example is the doctrine of the infallibility of a person's thoughts about his own present thoughts. According to this doctrine, the following is a necessary truth:

Infallibility. If x is thinking that x is thinking p, then x is thinking p.

To test whether the inverted relation thinking* satisfies this principle, let us substitute 'thinking*', for 'thinking':

If x is thinking* that x is thinking* p, then x is thinking* p.

Next let us expand the occurrences of 'thinking*' in accordance with the definition.[30] We obtain:

If x is thinking t(that x is thinking $t(p)$), then x is thinking $t(p)$.

To simplify this further, let us determine the identity of t(that x is thinking $t(p)$). Suppose that t is the particular/universal transformation. Consider an analogy: the proposition that $x < f(y)$. This proposition has no quantificational structure; i.e., it is not an existential generalization,[31] nor is it the result of predicating the property of being manifested of some universal. Therefore, t does not alter this proposition at all: t(that $x < f(y)$) = that $x < f(y)$. Now for exactly the same reason, t does not alter the proposition that x is thinking $t(p)$. That is, because this proposition has no quantificational structure, t(that x is thinking $t(p)$) = that x is thinking $t(p)$. Substituting this identity in the above principle, we obtain:

If x is thinking that x is thinking $t(p)$, then x is thinking $t(p)$.

Now suppose that p is some proposition with a complex quantificational structure; then $p \neq t(p)$. Let $q = t(p)$. Is the proposition that x is thinking $t(p)$ identical to the proposition that x is thinking q? Not at all. In fact, this is what we learned in our first lesson on the failure of substitutivity of co-referential expressions in intensional contexts. For example, the proposition that $7 <$ the number of planets is not identical to the proposition that $7 < 9$. Since 'the proposition that $7 <$ the number of planets' generates an intensional context and since 'the number of planets' is a descriptive phrase having narrow scope, it cannot be replaced by the co-referential expression '7' without changing the reference of the whole expression. Analogously, since 'the proposition that x is thinking $t(p)$' generates an intensional context and since '$t(p)$' is a descriptive expression having narrow scope (recall n. 16), it cannot be replaced by the co-referential expression 'q' without changing the reference of the whole expression. Thus, the proposition that x is thinking $t(p) \neq$ the proposition that x is thinking q. And, therefore, it is possible to think one and not the other. It is exactly this possibility

that creates a host of counterexamples to the last principle, proving that the inverted relation thinking* does not satisfy the original Infallibility principle.

Here is one such counterexample. Let $p =$ the proposition that the property of hitting something is manifest, and let $q =$ the proposition that something is such that the property of being hit by it is manifest, then $q = t(p)$. (Recall n. 25.) Suppose that x is not thinking q, i.e., that x is not thinking $t(p)$. Suppose, however, that x is thinking that something hits something, and suppose that x therefore thinks that he is thinking that something hits something. And suppose, finally, that, given the general human susceptibility to occasional logical errors, x mistakenly thinks that t(that hitting something is manifest) = that something hits something. Then, x might easily infer that he is thinking t(that hitting something is manifest). The resulting situation would be this: x would be thinking that x is thinking $t(p)$ and yet x would not be thinking $t(p)$. Hence, we have a counterexample to the principle that arises from the Infallibility principle when 'thinking*' is substituted for 'thinking'. Therefore, thinking* does not satisfy the Infallibility principle.

Moreover, this kind of breakdown is not an isolated phenomenon. Quinean constructions suffer analogous breakdowns in connection with a large range of principles concerning the mind's access to the contents of its own conscious intentional states. The following are two further examples, each of which has had an important role in the history of epistemology and philosophical psychology:

Privileged Access. If x is thinking p and x is entertaining the proposition that x is thinking p, then x knows that x is thinking p.

Introspection. If x is thinking p and x is entertaining the proposition that x is thinking p, then x introspects that x is thinking p.[32]

Using the previous example as a guide, one can easily show that the inverted relations thinking*, etc. do not satisfy these principles. And this conclusion holds regardless of which familiar Quinean transformation is used to define the inverted relations.

So far I have been discussing obstacles to our Quinean construction that arise in connection with the mind's access to the *contents* of its own conscious intentional states. I should now say a few words about obstacles that arise in connection with the mind's access to its own conscious *operations*. Consider the above Introspection principle. Let us follow our usual test procedure by putting in 'think*' for 'think', 'entertain*' for 'entertain', and 'introspect*' for 'introspect':

If x is thinking* p and x is entertaining* the proposition that x is thinking* p, then x introspects* that x is thinking* p.

Expanding 'entertain*', 'introspect*', and the unembedded occurrence of 'think*', we obtain:

If x is thinking $t(p)$ and x is entertaining the proposition t(that x is thinking* p), then x introspects t(that x is thinking* p).

Suppose for illustrative purposes that t is the particular/universal transformation.

Then, by considerations like those above, t(that x is thinking* p) = that x is thinking* p. Plugging in this identity, we obtain:

> If x is thinking $t(p)$ and x is entertaining the proposition that x is thinking* p, then x introspects that x is thinking* p.

This principle is plainly unacceptable for the sort of reasons discussed above. But there is a new defect as well. Let $q = t(p)$. Suppose x knows by introspection that he is thinking q. And suppose he knows the logical truth that he is thinking q if and only if he is thinking* p. In this situation, x could know by *inference* that he is thinking* p. However, that he is thinking* p is *in principle* simply not the sort of thing x (or anyone) can *directly* introspect; it is not the sort of thing someone can be aware of directly without the aid of some form of inference. The reason is that thinking* is not a conscious operation of mind, and the only relational propositions a person can directly introspect are those comprised of a conscious operation of mind and its immediate object. [For the same reason, a person x cannot directly introspect that x is thinking $t(p)$. True, x can directly introspect that x is thinking q, where $q = t(p)$. But, as we have seen, even though the descriptive proposition that x is thinking $t(p)$ and the purely relational proposition that x is thinking q are necessarily equivalent, they are quite distinct. Though x can know the descriptive proposition by inference via his introspective knowledge that he is thinking q and his logical knowledge that $q = t(p)$, it is impossible for him to introspect it. Only the purely relational proposition is a candidate for direct introspection.] The point, then, is that the range of the introspecting relation is necessarily restricted in something like the following way:

> x can introspect that xRy iff R is a conscious-operation-of-mind and it is possible that xRy.

That is, x can introspect that xRy if and only if R is either sensing, thinking, desiring, intending,..., or introspecting itself and it is possible that xRy. Thus, introspection defines a "hermetically sealed" circle of mutually tuned relations: all and only genuine conscious operations of mind are permitted entry. No system of relations thinking*, desiring*,..., introspecting* generated by our Quinean procedure can have this hermetically sealed character. On all such constructions, it is *impossible* for someone to introspect* that he is thinking* p. And on many of these constructions (e.g., the one based on the particular/universal transformation) it is *possible* for someone to introspect* propositions involving relations from outside the circle thinking*, desiring*,..., introspecting*; e.g., it is possible for someone to introspect* that he is thinking p.

A closely related theory about the interdependent structure of the conscious operations of mind is found in traditional empiricist psychology of experience (such as the theory propounded by John Locke). According to this theory, a person can in acts of reflection actually experience his own conscious operations of mind. So, for example, it is possible to experience thinking, desiring, deciding, introspecting, etc., and, indeed, it is possible to experience experiencing itself. However, according to the theory, it would be impossible to experience such things as thinking*, desiring*, etc. just because they are not conscious operations of mind. Our system of inverted relations clearly cannot satisfy this traditional empiricist theory.[33] What we need is an entirely new system of relations, one that mimics the behavior of the standard psychological relations "all they way down," even in their roles as constituents in the contents of conscious intentional states and as immediate objects of psychological relations themselves.

Of course, the foregoing are only examples of the kinds of principles for the self-conscious rational mind that we should take into account; there are many more. One especially worth mentioning arises in connection with the fact that rational beings can know a priori a great number of necessary truths. We may approximate this fact with the following principle:

> *A Priori Knowledge*. If p is necessary and x clearly and distinctly understands p, then x knows p.[34]

(To understand a proposition clearly and distinctly is, roughly, to "grasp its full significance." This is far stronger than merely entertaining the proposition.) Since this general principle of a priori knowledge is elementary, our inverted relations knowing* and clear and distinct understanding* do satisfy it. However, there are infinitely many specific principles of a priori knowledge that are not satisfied by the inverted relations, namely, non-elementary knowledge principles concerning specific logical relationships holding between the standard psychological relations and their inverted counterparts.[35]

Now let P be our best theory for the self-conscious rational mind. Suppose P includes the sorts of principles we have been discussing. This makes it much harder for a system of relations to satisfy P. Therefore, it is much more likely that P will be satisfied by only one system of relations–namely, the standard psychological relations–and, in turn, it is much more likely that P will serve as the basis for adequate functional definitions of the standard psychological relations. Now the principles that we have considered in this section might be controversial. However, to refute functionalism once and for all, either we must exclude all such principles or we must construct a new system of inverted relations that satisfies P. It is out of the question to dispute all such principles; surely at least some versions of some of these principles are sound. Therefore, one must adopt the second strategy, i.e., one must give a construction that works even when P contains such principles.

The difficulty of giving such a construction is magnified because we must allow for the possibility that P attributes a very strong modal value (indeed, even metaphysical necessity) to each of its constituent principles, including those of the sorts we have considered in this section. As I indicated earlier, none of the currently popular criticisms of functionalism comes close to providing a construction that fills this requirement. The goal of this section is to give such a construction. In effect, the construction

is a global generalization of the construction given in the previous section. Specifically, I will define a new system of inverted relations that, like the earlier one, may be thought of as comprising an "anti-mind." This time, however, even the totality of principles for the self-conscious rational mind will be blind to the distinction between anti-mind and the true mind.

Before I launch into the construction it might be helpful for me to say something about the logical form of the kind of sentences that are used to express non-elementary principles for the propositional attitudes. Recall our first example, 'Whenever someone thinks that he thinks something, then he in fact thinks it'. On its intended reading, this sentence imposes the following condition on the thinking relation T: for any x and p, if x stands in relation T to the result of applying T to x and p, then x stands in relation T to p. Thus, a canonical representation of (this reading of) the sentence would be something like this:

$$(\forall x)\,(\forall p)\,(xT[xTp] \rightarrow xTp)$$

where all *three* occurrences of 'T' express the same relation, namely, thinking. Only then will we capture the condition imposed on the thinking relation by the original (reading of the) sentence.

To appreciate better the significance of this point, notice that a strict "language-of-thought" theory allows for no natural way to represent the original (reading of the) sentence or others like it. For on a strict "language-of-thought" theory, pure syntactic entities are the only objects of thought. Therefore, the closest one could come to the original sentence would be something like this:

$$(\forall x)\,(\forall p)\,(xT'xTp' \rightarrow xTp).$$

This fails to capture the condition imposed on thinking by the original sentence, for the second occurrence of 'T' does not stand for the thinking relation but rather for the symbol 'T' itself. (Moreover, the second occurrences of 'x' and 'p' do not serve as variables ranging over persons and objects of thought but rather as names for the symbols 'x' and 'p' themselves.) There are certain rather artificial techniques for avoiding this outcome. But each of these methods, in effect, permits selected non-linguistic entities–e.g., real psychological relations such as the thinking relation–right into the objects of thought. Now anti-representational realists may accept this feature, but strict language-of-thought theorists may not. For this feature violates the basic philosophical tenet of their position, namely, that every constituent in an object of thought is a mere representation, i.e., an expression in the language of thought. But psychological relations are not expressions in the language of thought; rather they are what relate people to expressions in the language of thought (or so the language-of-thought theory goes)![36]

I begin the construction with an informal characterization. The construction is based on two intuitive ideas, the first of which is this. Suppose that, for every standard psychological relation, there is a distinct relation (called the "counterpart") that is

necessarily like the original in two significant ways. I will illustrate with the example of thinking and its counterpart, which I will call thinking*. (This relation should not be confused with the one constructed in the previous section.) First, the extensions of thinking and thinking* are necessarily the same when it comes to propositions that do not "involve" the standard mental relations or their counterparts. So, for example, if q is a proposition concerning only individuals, necessarily, x thinks q iff x thinks* q. Second, when it comes to propositions that do "involve" mental relations or their counterparts, the extensions of the two relations are connected as follows: necessarily, for any such proposition p, x thinks p iff x thinks* p^*, where p^* arises from p by "substituting" mental relations for their counterparts and counterparts of mental relations for their originals wherever they "occur" in p. (For example, if p is the proposition that x thinks something, then p^* is the proposition that x thinks* something. Thus, necessarily, x thinks that x thinks something iff x thinks* that x thinks* something.) Now given that these relationships hold necessarily, thinking and thinking* will, I submit, be functionally indistinguishable. Nevertheless, they will be significantly different. After all, the type of proposition that is the characteristic object of the thinking relation is very "fine-grained." Therefore, if p and p^* are as specified in the example, p and p^* must be distinct, for p "contains" thinking where p^* "contains" thinking*. Yet for any two distinct propositions, it is at least possible that someone x thinks one and not the other. Therefore, suppose that x thinks p, then it is at least possible that x does not at that moment also think p^*. However, necessarily, x thinks p iff x thinks* p^*. Hence, it is at least possible that thinking and thinking* are not co-extensive, and this constitutes a significant difference between them. Thus, the mere distinctness of two relations (such as thinking and thinking*) that are otherwise indistinguishable can actually generate a distinction in their extensions when it comes to higher-level propositions that "contain" them as ' constituents." This fact about the "fine-grained" intensionality of the standard propositional attitudes leads directly to model-theoretic proofs of the following: the standard psychological relations cannot be implicitly defined by any of a wide variety of psychological theories.[37]

My goal, however, is to *define explicitly* a system of new relations that comprises an anti-mind. To carry out this construction, we must incorporate our second intuitive idea. The function of this idea will be to guarantee that, when defined, our new relations thinking*, desiring*, etc. truly are distinct from the standard psychological relations. The idea is once again that of a Quinean "inversion" of our conceptual scheme, except that this time the inversion will be more radical. Recall that in our first construction the Quinean transformation t leaves intact every embedded proposition that is a constituent of the proposition upon which t is operating. By contrast, in its new, more radical role, t would, in effect, be applied over and over again to transform every embedded proposition "all the way down." Thus, in the new construction all the following will hold:

x thinks* that rabbithood is manifest iff x thinks that there exists a rabbit.

x thinks* that x thinks* that rabbithood is manifest iff x thinks that x thinks that there exists a rabbit.

x thinks* that x thinks* that x thinks* that rabbithood is manifest iff x thinks that x thinks that x thinks that there exists a rabbit.

. . . .

Observe that t is applied to every embedded proposition "all the way down" and also that every occurrence of thinking* is replaced by an occurrence of thinking. The aim of the new construction is to generalize this inversion procedure. That is, we will apply t "all the way down" to all fine-grained intensions–concepts as well as propositions. Moreover, for every property or relation F, we will define an inverted counterpart F^* such that F^* will play F's role and F will play F^*'s role at every occurrence in every fine-grained intension. (Of course, F and F^* will be identical in certain key cases, namely, where F is a physical property or relation, or F is one of the familiar logical relations such as identity or necessary equivalence.) The way this generalized inversion procedure will be achieved is by the definition of an operation $*$ which turns every property or relation F into an inverted counterpart F^* and which turns F^* right back into the original property or relation F. That is, $*(*(F)) = F$.

In defining this operation, we will use some relatively powerful logical machinery. Let me explain why. If we were dealing with specific versions of functionalism that invoke only some kinds of non-elementary psychological principles, a less powerful construction might suffice. For example, consider a version of functionalism that invokes only non-elementary principles of the kind that can be expressed within a strict theory of types (such as th..t of Russell or of Church); we could then construct an operation $*$ by straightforward inductive means. However, the kinds of non-elementary principles we have been discussing in this section require a type-free logical setting so that a psychological relation (e.g., thinking) can occur as a constituent of propositions that fall within its very own range. This special form of impredicativity appears to undermine an inductive approach to the definition of $*$ and of the $*$-counterparts of the standard psychological relations. There are two reasons why this is so. The first is this. Suppose that we have a candidate inductive characterization of some relation. Then, we can transform such a characterization into a direct definition of the relation simply by taking either the intersection or the union of all relations that satisfy the characterization. However, for any relation that exhibits the special kind of impredicativity that the $*$-relations must exhibit, these familiar methods appear to be of no use. For, if there should happen to exist more than one relation satisfying the inductive characterization of, e.g., thinking*, then, given the special impredicativity these relations would have, their intersection would be too "small" and their union would be too "large." And there appears to be no reason why there could not in fact exist *several* distinct relations satisfying any candidate inductive characterization of thinking*. The second problem, which is really far more serious, is that an inductive approach seems unable to provide any guarantee that there exists even *one* relation

satisfying an inductive characterization of, say, thinking*. When we try to use an inductive rationale to prove the existence of such a relation, we seem simply to go round in a circle: given its special impredicativity, we must in effect already know that thinking* exists in order to be in a position to prove its existence by building up its extension in inductive stages. It appears, therefore, that something more powerful than an inductive construction is needed to break out of such circles. What I will show is that a new kind of diagonalization does the job. Moreover, this diagonalization, which is based upon the construction of a special kind of substitution function, will lead to a *direct* definition of the operation $*$ and, in turn, to direct definitions of all the $*$-counterparts of the standard psychological relations.

The construction is formulated within the framework of informal intensional logic.[38] However, many of the ideas used are borrowed from the formal intensional logic developed in *Quality and Concept*. (This is a convenience; there are alternate, but more cumbersome, terms in which to formulate the construction.) According to this formal theory, each fine-grained intension–i.e., each concept and each proposition–has a unique logical form that is determined by decomposing the intension under the inverses of certain fundamental logical operations (called *proposition-forming operations*), e.g., existential generalization, negation, conjunction, singular predication, relativized predication, etc. I call the outcome of this type of decomposition procedure the *decomposition tree* for the original fine-grained intension. Some of the nodes in a decomposition tree may be terminal; others are not terminal. A *terminal node* is one occupied by an entity that cannot be decomposed further under the inverses of the fundamental proposition-forming operations. Such entities, then, are simple with respect to logical form. In each decomposition tree, a node can play one of two roles–formal or substantive. For each application of the inverse of one of the fundamental proposition-forming operations, there is a corresponding node occupied by that operation. Such nodes are called *formal nodes*. All others are called *substantive*. Among the substantive nodes there is a special kind called *subject nodes*. These nodes are distinguished by the fact that the entities occupying them are the subjects of associated predications. (For example, a occupies a subject node in the decomposition tree of the proposition $[Fa]$ since this proposition results from a singular predication of the property F of the item a.) Every substantive node that is not a subject node is called a *predicate node*. (Thus, the property F occupies a predicate node in the decomposition trees of the propositions $[Fa]$, $[(\exists x)Fx]$, etc.)

I will call a fine-grained intension *purely particular* if its decomposition tree contains no predicate node occupied by the property of being manifested. A fine-grained intension is then called *pure* if its decomposition tree can be obtained from that of some purely particular intension by the following procedure. Working up from terminal nodes, we may whenever we wish replace an existential generalization with an associated predication whose predicate node is occupied by the property of being manifest, but if we do then all the nodes from which the altered one descends must be adjusted accordingly. For example, if M is the property of being manifest, then

$[(\exists x)(\exists y)Rxy]$, $[(\exists x)M[Rxy]_y]$, $[M[(\exists y)Rxy]_x]$, and $[M[M[Rxy]_y]_x]$ are the pure propositions that arise via this procedure from the purely particular proposition $[(\exists x)(\exists y)Rxy]$. The intuitive idea here is this. A fine-grained intension is pure if each predicative occurrence of M in its decomposition tree is, in effect, a "quasi-formal" node. Specifically, taken together with its adjacent predication operation, each such occurrence of M behaves much the same way as a genuine formal node occupied by the operation of existential generalization. With this in mind, I will as a terminological convenience extend my use of the term 'formal node' so that, whenever the property M occupies a predicate node in the decomposition tree of a pure fine-grained intension, I will also call that node a *formal node*. No confusion should result from this liberalization.

We come now to the generalized Quinean transformation T, which is like the particular/universal transformation t used earlier except that it transforms fine-grained intensions "all the way down." Stated in intuitive terms, the definition of T is this. If x is not a pure fine-grained intension, $T(x) =_{df} x$, and if x is a pure fine-grained intension, then $T(x) =_{df}$ the pure fine-grained intension whose decomposition tree can be obtained from that of x by the following procedure. Working up from the terminal nodes in x's tree, (1) we replace every existential generalization with a suitably associated predication whose predicate node is occupied by M, (2) we replace every predication whose predicate node is occupied by M with an associated existential generalization, and (3) we make corresponding adjustments at every node from which any altered node descends. So, for example:

$T([(\exists x)(\exists y)Rxy]) = [M[M[Rxy]_y]_x]$
$T([M[M[Rxy]_y]_x]) = [(\exists x)(\exists y)Rxy]$
$T([(\exists x)M[Rxy]_y]) = [M[(\exists y)Rxy]_x]$
$T([M[(\exists y)Rxy]_x]) = [(\exists x)M[Rxy]_y]$
$T([(\exists u)Ru[(\exists x)(\exists y)Rxy]]) = [M[Ru[M[M[Rxy]_y]_x]]_u]$.

Now let us extend T to finite sequences. If β is a sequence $\beta_1,\ldots,\beta_n$ then we define $T(\beta)$ to be the sequence β' that is just like β except that, for any fine-grained intension β_i that is an element in β, $T(\beta_i)$ is the corresponding element in β'.

So far I have been discussing fine-grained intensions–concepts and propositions. I must now say a word about "coarse-grained" intensions–i.e., properties, relations, and conditions. These intensions are coarse-grained in the sense that their identity conditions are less strict than those of fine-grained intensions. Specifically, coarse-grained intensions are identical if they are necessarily equivalent. Now just as fine-grained intensions are the values of the fundamental proposition-forming operations, coarse-grained intensions are values of fundamental logical operations of a corresponding type; I call these *condition-forming operations*. Of course, since coarse-grained intensions have much weaker identity conditions, the inverses of the condition-forming operations never determine unique decomposition trees for any coarse-grained intension; indeed, each coarse-grained intension has infinitely many

such trees. These trees, however, have nothing to do with logical form since coarse-grained intensions are simple with respect to logical form.

In light of the distinction between condition-forming and proposition-forming operations, a terminological stipulation is now in order. Henceforth, when I speak of the singular predication operation, I will always mean the condition-forming operation of singular predication; I will never mean the proposition-forming operation. (Both operations are discussed at length in *Quality and Concept*.) In this connection a few examples of how this operation works might be helpful. The condition that I think, for example, is the result of predicating of me the property thinking; symbolically, [I think] = Pred($[x$ thinks$]_x$,me). Consider next the property of loving Mary. This is the result of predicating of Mary the relation loving; symbolically, $[x$ loves Mary$]_x$ = Pred($[x$ loves $y]_{xy}$,Mary). Consider, finally, an example of self-predication: the property of being identical to the identity relation. This property is the result of predicating the identity relation of the identity relation itself; symbolically, $[x = [x = y]_{xy}]_x$ = Pred($[x = y]_{xy}$,$[x = y]_{xy}$).[39]

Next a word on notation is needed. In what follows (and in the previous paragraph) I use the bracket notation to denote coarse-grained intensions. Specifically, I use $[A]$ to denote the condition that A; $[A(v)]_v$ to denote the property of being a v such that $A(v)$; and $[A(v_1,\ldots,v_n)]_{v_1\ldots v_n}$ to denote the relation among $v_1,\ldots,v_n$ such that $A(v_1,\ldots,v_n)$. I will not use this notation for denoting fine-grained intensions (concepts or propositions). For the sake of readability I will use lower-case Greek letters α, β, γ, δ as variables that range over finite sequences. These same letters numerically subscripted will be used as variables that range over the relevant elements of the associated sequences; thus, α_i will indicate the i-th element of the sequence α. (I identify finite sequences, not with set-theoretic entities, but with appropriate properties.)[40] Notice that, although for readability I use several sorts of variables, the definition of $*$ and the ensuing theorems about it will be set forth so that they can be easily written out in a first-order language with a single sort of variable. This is important philosophically, for it guarantees that the construction will have maximum generality.[41]

Let c be any function, and let α be any sequence $\alpha_1,\ldots,\alpha_k$ for $k \geq 1$. I define a substitution function $\alpha(G/c(G))$ on sequences α as follows. Suppose that c always maps properties to properties and i-ary relations to i-ary relations, for each $i \geq 2$. Then $\alpha(G/c(G))$ is the new sequence β whose elements $\beta_1,\ldots,\beta_k$ arise from those of α as a result of the following substitutions. There are three cases. First, an element α_i in α might be an individual or condition. In this case, $\beta_i = \alpha_i$. Second, α_i might be a property or relation. In this case, if $\alpha_i = c(c(\alpha_i))$, then $\beta_i = c(\alpha_i)$, but if $\alpha_i \neq c(c(\alpha_i))$, then $\beta_i = \alpha_i$. Third, α_i might be a fine-grained intension. In this case, β_i is the fine-grained intension whose decomposition tree is just like that of α_i except for the following. Consider any property or relation G that occurs at a nonformal terminal node n in α_i's decomposition tree. If $G = c(c(G))$, then $c(G)$ occurs at node n in β_i's decomposition tree (and corresponding adjustments are then made at all nodes from which n descends), but if $G \neq c(c(G))$, then G itself occurs at node n in β_i's decomposition

tree. On the other hand, suppose c does not always map properties to properties and i-ary relations to i-ary relations, for each $i \geq 2$; then $\alpha(G/c(G))$ is just α itself.

Consider an example. Suppose that c always maps properties to properties and i-ary relations to i-ary relations, for each $i \geq 2$. Let the sequence α consist of three elements: an individual x, the relation believing, and the proposition that someone believes something. Finally, let believing* = c(believing). Then, if believing = $c(c$(believing)), $\alpha(G/c(G))$ consists of: x, believing*, and the proposition that someone believes* something. On the other hand, if believing $\neq c(c$(believing)), then $\alpha(G/c(G))$ is simply α itself. Of course, if c does not always map properties to properties and i-ary relations to i-ary relations, for each $i \geq 2$, then $\alpha(G/c(G))$ is once again α itself.

I next introduce an operation d on properties and relations. Again there are three cases. First, for $i \geq 2$, if F is an i-ary relation that is not necessarily null, then $d(F) =_{df} [\alpha \text{ satisfies } F]_{\alpha}$, i.e., the property of being a sequence α that satisfies F. Second, if F is a property that is not necessarily null, then $d(F) =_{df} [(\exists x)\,(x \text{ has } F \;\&\; v = [u = x]_u)]_v$, i.e., the property of being identical to the property of being identical to some item x that has F. Finally, in the degenerate case where F is a necessarily null i-ary relation, $i \geq 1$, $d(F) =_{df} [v = i]_v$, i.e., the property of being the natural number i, which is F's degree. Since d is a one-one function, we may define a related one-one function e which takes any property in the range of d back to the property or relation that this property came from: $e(F) =_{df} d^{-1}(F)$ if F is in the range of d; otherwise $e(F) =_{df} F$. The effect of d is to provide a unique property that acts as a "code" for any given property or relation, and the effect of e is to "decode" any such property by turning it back into the original uncoded property or relation it came from.

Let S be some binary relation. Then, given the above definitions, the complex S-based substitution function $\lambda(G/e(\text{Pred}(S,G)))$ is defined on all sequences α, and hence, so is the substitution function $T(\alpha(G/e(\text{Pred}(S,G))))$. Let us call the latter the *S-inversion of* α. Using this notion, I will now define 'β arises from an F-satisfier via S-inversion' for all properties and relations F. Once again there are three cases. First, for $i \geq 2$, let F be an i-ary relation that is not necessarily null. Suppose that, for some α, α satisfies F and β is the sequence that arises from α via S-inversion, i.e., suppose that

$$(\exists\alpha)\,(\alpha \text{ satisfies } F \;\&\; \beta = T(\alpha(G/e(\text{Pred}(S,G))))).$$

Then, I will say that β *arises from an F-satisfier via S-inversion*. Second, let F be a property that is not necessarily null. Suppose that, for some 1-ary sequence α, α has F and β is the property of being identical to the item that arises from α via S-inversion, i.e., suppose that

$$(\exists\alpha)\,(\alpha \text{ has } F \;\&\; \beta = [v = T(\alpha(G/e(\text{Pred}(S,G))))]_v).$$

Then, I will again say that β *arises from an F-satisfier via S-inversion*. Third, for the degenerate case, let F be a necessarily null i-ary relation, $i \geq 1$, and let $\beta = i$. Then,

for convenience, I will again say that β *arises from an F-satisfier via S-inversion*.

Consider the following relation Q:

$$[(\forall S)\,(\text{Pred}(R,R) = S \rightarrow \beta \text{ arises from an } F\text{-satisfier via } S\text{-inversion})]_{\beta FR}.$$

Q holds among sequences β, properties or relations F, and relations R such that, if S is the result of predicating R of R, then β arises from an F-satisfier via S-inversion. Now for the diagonalization. Let Q be predicated of Q. The result is:

$$[(\forall S)\,(\text{Pred}(Q,Q) = S \rightarrow \beta \text{ arises from an } F\text{-satisfier via } S\text{-inversion})]_{\beta F}.$$

Let us abbreviate this as follows: $[\text{S}(\beta,F)]_{\beta F}$. Then the following is our definition of the operation $*$ on properties and relations F:

$$*(F) =_{df} e([\text{S}(\beta,F)]_{\beta}).$$

In the coming paragraphs I will show that this operation is in fact the one we are seeking; that is, I will show that it meets all the requirements set forth in our informal characterization of $*$.

The first step is to prove the following theorem for all sequences β and all properties and relations F:

$$\beta \text{ satisfies } *(F) \text{ iff } (\exists\alpha)\,(\alpha \text{ satisfies } F \;\&\; \beta = T(\alpha(G/*(G)))).$$

There are three cases. The first case is that in which F is an i-ary relation, $i \geq 2$, that is not necessarily null. For such relations F the theorem follows by putting together a chain of six biconditionals, which I will now present. By the definition of $*$:

(1) β satisfies $*(F)$ iff β satisfies $e([\text{S}(\beta,F)]_{\beta})$.

Since F is an i-ary relation that is not necessarily null, we can easily check that $[\text{S}(\beta,F)]_{\beta}$ is a non-necessarily-null property of i-ary sequences β. Therefore, by the definition of e,

(2) β satisfies $e([\text{S}(\beta,F)]_{\beta})$ iff β has the property $[\text{S}(\beta,F)]_{\beta}$.

The abstraction principle governing the property $[\text{S}(\beta,F)]_{\beta}$ tells us:

(3) β has the property $[\text{S}(\beta,F)]_{\beta}$ iff $\text{S}(\beta,F)$.

By expanding 'S' in accordance with the definition, we obtain:

(4) $\text{S}(\beta,F)$ iff
$(\forall S)\,(\text{Pred}(Q,Q) = S \rightarrow \beta$ arises from an F-satisfier via S-inversion).

Since F is an i-ary relation, $i \geq 2$, that is not necessarily null, it follows by the definition of 'β arises from an F-satisfier via S-inversion' that:

(5) $(\forall S)\,(\text{Pred}(Q,Q) = S \rightarrow \beta$ arises from an F-satisfier via S-inversion) iff
$(\forall S)(\text{Pred}(Q,Q) = S \rightarrow (\exists\alpha)\,(\alpha \text{ satisfies } F \;\&\; \beta = T(\alpha(G/e(\text{Pred}(S,G))))))$.

Our sixth biconditional requires a more extended argument. By its definition, $[S(\beta,F)]_{\beta F}$ is the unique outcome of predicating Q of Q. But notice that the condition imposed on S in the antecedent of the expanded version of 'S' is that $\text{Pred}(Q,Q) = S$; i.e., S must be the unique outcome of predicating Q of Q. Therefore, $S = [S(\beta,F)]_{\beta F}$. At the same time, $\text{Pred}([S(\beta,F)]_{\beta F},G) = [S(\beta,G)]_{\beta}$. It follows that $e(\text{Pred}(S,G)) = e([S(\beta,G)]_{\beta})$. But by the definition of $*$, $*(G) = e([S(\beta,G)]_{\beta})$. Hence, $e(\text{Pred}(S,G)) = *(G)$. Thus, $T(\alpha(G/e(\text{Pred}(S,G)))) = T(\alpha(G/*(G)))$. Applying this identity, we obtain our sixth biconditional:

(6) $(\forall S)\,(\text{Pred}(Q,Q) = S \rightarrow (\exists\alpha)\,(\alpha \text{ satisfies } F \,\&\, \beta = T(\alpha(G/e(\text{Pred}(S,G))))))$ iff $(\exists\alpha)\,(\alpha \text{ satisfies } F \,\&\, \beta = T(\alpha(G/*(G))))$.

Putting these six biconditionals together, we directly establish our theorem for the case in which F is an i-ary relation, $i \geq 2$, that is not necessarily null. By an analogous argument we can establish the theorem for the case in which F is a property that is not necessarily null. And the theorem holds trivially for the case in which F is a necessarily null property or relation. Thus, the theorem holds for all properties and relations F.

What the theorem says is this. For any sequence β, β satisfies $*(F)$ if and only if, for some sequence α, α satisfies F and β is the outcome of performing the compound substitution $T(\alpha(G/*(G)))$. Hence, for any sequence α, if α satisfies F, then $T(\alpha(G/*(G)))$ satisfies $*(F)$. Now let us assume that the compound substitution $T(\alpha(G/*(G)))$ is one-one. That is, assume that, for all α and δ, $\alpha = \delta$ if $T(\alpha(G/*(G))) = T(\delta(G/*(G)))$. Given this assumption, it also follows that, for all sequences α, if $T(\alpha(G/*(G)))$ satisfies $*(F)$, then α satisfies F. Putting these two conclusions together, we obtain:

$T(\alpha(G/*(G)))$ satisfies $*(F)$ iff α satisfies F

for all sequences α.

Let us assume that $*$ always maps properties to properties and i-ary relations to i-ary relations, for each $i \geq 2$. And let us also assume that $G = *(*(G))$ for all properties and relations G. Then, given the conclusion just reached, the sequences $T(\alpha(G/*(G)))$ that satisfy $*(F)$ will be precisely those stipulated in our original informal characterization of the extension of $*(F)$. For they will be the sequences we obtain from those satisfying F if we perform the compound substitutions stipulated in our informal characterization. Specifically, if an element in α is a property or relation G, then the corresponding element in the new sequence is $*(G)$. And if an element in α is a fine-grained intension, then the corresponding element in the new sequence is the fine-grained intension arising from the original via the following two-step procedure. First, we form an intermediate fine-grained intension whose decomposition tree is like that of the original intension except that, for each property or relation G occupying a nonformal terminal node in the original tree, $*(G)$ takes G's place in the new tree, and, in turn, all appropriate adjustments reflecting these changes are made in the new tree. Second, this intermediate intension is then subjected to the generalized Quinean transformation T.

Therefore, if we can prove our three assumptions–(1) the operation $*$ always maps properties to properties and i-ary relations to i-ary relations, for each $i \geq 2$; (2) the compound substitution $T(\alpha(G/*(G)))$ is one-one; and (3) $G = *(*(G))$ for all G–then we will be certain that we have indeed constructed the operation $*$ we have been seeking.

To prove proposition (1), assume its contrary. Then, by the definition of $\alpha(G/*(G))$, we know that $\alpha(G/*(G)) = \alpha$. Given this and the above theorem it is trivial to verify that $*$ always maps properties to properties and i-ary relations to i-ary relations, for each $i \geq 2$. Thus, the hypothesis leads to a contradiction, and so proposition (1) must be true.

The generalized Quinean transformation T was expressly constructed as a one-one operation. Thus, to prove proposition (2), we must only prove that $\alpha = \delta$ if $\alpha(G/*(G)) = \delta(G/*(G))$. Suppose that $\alpha(G/*(G)) = \delta(G/*(G)) = \beta$. To show that $\alpha = \delta$, we show that $\alpha_i = \delta_i$ for each element α_i in α and δ_i in δ. Given proposition (1), it is easy to verify that α_i and β_i must be the same type of entity–individual, condition, property, relation, or fine-grained intension. Likewise for δ_i and β_i. Therefore, α_i and δ_i must be the same type of entity, too. Suppose α_i and δ_i are individuals or conditions. Then since the substitution procedure leaves these types of entities unchanged, $\alpha_i = \beta_i$ and $\delta_i = \beta_i$. So $\alpha_i = \delta_i$. Suppose that α_i and δ_i are properties or that they are relations. First, we show that $*(*(\alpha_i)) = \alpha_i$ iff $*(*(\delta_i)) = \delta_i$. Suppose otherwise. For example, suppose $*(*(\alpha_i)) \neq \alpha_i$ and $*(*(\delta_i)) = \delta_i$. Then, given proposition (1) and the definition of $\alpha(G/*(G))$, it follows that $\beta_i = \alpha_i$. And given proposition (1) and the definition of $\delta(G/*(G))$, it follows that $\beta_i = *(\delta_i)$. Given these two conclusions, it follows that $\alpha_i = *(\delta_i)$. Hence, by an application of $*$, $*(\alpha_i) = *(*(\delta_i))$. Then, by our hypothesis that $*(*(\delta_i)) = \delta_i$, it follows that $*(\alpha_i) = \delta_i$. Therefore, by another application of $*$, $*(*(\alpha_i)) = *(\delta_i)$. Thus, by our conclusion that $\alpha_i = *(\delta_i)$, it follows that $*(*(\alpha_i)) = \alpha_i$ contradicting our hypothesis that $*(*(\alpha_i)) \neq \alpha_i$. By an analogous argument we can show that it is impossible that $*(*(\alpha_i)) = \alpha_i$ and $*(*(\delta_i)) \neq \delta_i$. Therefore, there are only two possible cases: (1) $*(*(\alpha_i)) \neq \alpha_i$ and $*(*(\delta_i)) \neq \delta_i$, and (2) $*(*(\alpha_i)) = \alpha_i$ and $*(*(\delta_i)) = \delta_i$. Suppose the former. Then, given proposition (1) and the definition of $\alpha(G/*(G))$, $\beta_i = \alpha_i$. And given proposition (1) and the definition of $\delta(G/*(G))$, $\beta_i = \delta_i$. Hence, $\alpha_i = \delta_i$. Or suppose the latter. Then, given proposition (1) and the definition of $\alpha(G/*(G))$, $\beta_i = *(\alpha_i)$. And given proposition (1) and the definition of $\delta(G/*(G))$, $\beta_i = *(\delta_i)$. Thus, $*(\alpha_i) = *(\delta_i)$. Hence, by an application of $*$, $*(*(\alpha_i)) = *(*(\delta_i))$. Therefore, by the hypothesis that $*(*(\alpha_i)) = \alpha_i$ and $*(*(\delta_i)) = \delta_i$ it follows that $\alpha_i = \delta_i$ once again. Consider, finally, the possibility that α_i and δ_i are fine-grained intensions. We know, by the definition of $*$, that any difference between the decomposition tree of α_i and that of β_i arises from differences in their nonformal terminal nodes resulting from replacing some property or relation G with $*(G)$. Likewise for the decomposition tree of δ_i and that of β_i. Therefore, any difference between the decomposition tree of α_i and that of β_i results from a difference between a property or relation G occupying a nonformal terminal node in the one and a property or relation H occu-

pying the corresponding nonformal terminal node in the other. However, by an argument exactly like that given a moment ago, we can prove that for every such G and H, $G = H$. Therefore, there can ɔe no difference between the decomposition tree of α_i and that of δ_i. Hence, $\alpha_i = \delta_i$ once again. Since this exhausts all possible cases, $\alpha = \delta$. And this completes our proof.

Next we prove that $*(*(F)) = F$ for any property or relation F. For each sequence α, there is a unique sequence β that arises from α via our substitution procedure and there is a unique sequence γ that arises from β via the same procedure; i.e., for each α, there is a unique β and a unique γ such that $\beta = T(\alpha(G/*(G)))$ and $\gamma = T(\beta(G/*(G)))$. We begin by showing that each element α_i in α is identical to the corresponding element γ_i in γ. Let an element α_i in α be an individual (or condition). Then the corresponding element β_i in β–and, in turn, the element γ_i in γ–is just that individual (condition). Thus, $\alpha_i = \gamma_i$. Let α_i be a property or a relation. Suppose that $\alpha_i \neq *(*(\alpha_i))$. Then, given proposition (1) and the definition of $\alpha(G/*(G))$, $\beta_i = \alpha_i$. Therefore, by substitution of β_i for α_i in the hypothesis, $\beta_i \neq *(*(\beta_i))$. Thus, given proposition (1) and the definition of $\beta(G/*(G))$, $\gamma_i = \beta_i$. Hence, by our conclusion that $\beta_i = \alpha_i$, $\alpha_i = \gamma_i$. On the other hand, suppose that $\alpha_i = *(*(\alpha_i))$. Then, given proposition (1) and the definition of $\alpha(G/*(G))$, $\beta_i = *(\alpha_i)$. Hence, by an application of $*$, $*(\beta_i) = *(*(\alpha_i))$. Therefore, by the hypothesis, $*(\beta_i) = \alpha_i$. Hence, by another application of $*$, $*(*(\beta_i)) = *(\alpha_i)$. So, by our conclusion that $\beta_i = *(\alpha_i)$, it follows that $*(*(\beta_i)) = \beta_i$. Therefore, given proposition (1) and the definition of $\beta(G/*(G))$, $\gamma_i = *(\beta_i)$. Hence, by our conclusion that $*(\beta_i) = \alpha_i$, $\gamma_i = \alpha_i$ once again. Finally, let α_i be a fine-grained intension. We know that there is a one-to-one correspondence between the nonformal terminal nodes in α_i's decomposition tree and those in β_i's decomposition tree and between those in β_i's decomposition tree and those in γ_i's decomposition tree. Hence, there is a one-to-one correspondence between the nonformal terminal nodes in α_i's decomposition tree and those in γ_i's. Moreover, our generalized Quinean transformatio· T is such that $\Theta = T(T(\Theta))$ for every fine-grained intension Θ. Therefore, the formal nodes in α_i's decomposition tree are the same as those in γ_i's decomposition tree. Hence, if there is a difference between α_i's decomposition tree and γ_i's decomposition tree, it must result from a difference between the items occupying the nonformal terminal nodes in α_i's tree and the items occupying the corresponding nonformal terminal nodes in γ_i's tree. However, individuals and conditions are always left alone by the substitution procedure. So if there is a difference between any item G occupying a nonformal terminal node in α_i's tree and the item H occupying the corresponding nonformal terminal node in γ_i's tree, then G and H are properties or they are relations. However, by an argument exactly like that given for properties and relations a moment ago, $G = H$ if G and H are properties or if they are relations. It follows, therefore, that there can be no difference between the decomposition trees of α_i and γ_i. Hence $\alpha_i = \gamma_i$. Since $\alpha_i = \gamma_i$ for all α_i in α and γ_i in γ, $\alpha = \gamma$. And, this holds for all α and γ such that, for some β, $\beta = T(\alpha(G/*(G)))$ and $\gamma = T(\beta(G/*(G)))$. Now given our first theorem and proposition (2), we know that for each such α, β, and γ, α satisfies F iff β satisfies $*(F)$ iff γ satisfies $*(*(F))$. Combining this with our conclusion that $\alpha = \gamma$ for each such α and γ, we obtain the following. For each α, α satisfies F iff α satisfies $*(*(F))$. Since all of the above holds necessarily, it follows, therefore, that F and $*(*(F))$ are necessarily equivalent. But since properties and relations are coarse-grained intensions, they are identical if necessarily equivalent. Thus, $F = *(*(F))$ for all properties and relations F. And this completes the proof.

As a notational convenience, let us extend our usage of $*$ to fine-grained intensions: if Θ is a fine-grained intension, $*(\Theta)$ is defined to be the fine-grained intension that results when the 1-ary sequence Θ is subjected to our substitution procedure for fine-grained intensions; i.e., $*(\Theta) =_{df} T(\Theta(G/*(G)))$. Now when we were dealing with the elementary belief/desire model in the previous section, our argument repeatedly invoked the following lemma: for any proposition p, $p \approx_N t(p)$; i.e., p is necessarily equivalent to $t(p)$. In the present setting, the lemma that would play an analogous role would be this: for any proposition p, $p \approx_N *(p)$.

To facilitate the proof of this lemma, let us also extend our usage of $*$ to cover individuals and conditions: if Θ is an individual or a condition, $*(\Theta)$ is defined to be Θ itself. Given this, $*$ is now defined on all types of things–individuals, properties, relations, conditions, and fine-grained intensions. The first step is to use proposition (3) to prove the following generalization: $\Theta = *(*(\Theta))$ for any item Θ. I will omit the proof since it is straightforward. Next, for any item Θ, if $*(\Theta) = \tau$, let us use the expression Θ^* to denote τ. Also, let us use $[A]$ to denote the proposition that A.

Now for the proof of the lemma. If α is a sequence $\alpha_1,\ldots,\alpha_i$, then by definition $T(\alpha(G/*(G)))$ is the sequence $\alpha_1^*,\ldots,\alpha_i^*$. We have proved that, necessarily, α satisfies F iff $T(\alpha(G/*(G)))$ satisfies F^*. Thus, necessarily, $\alpha_1,\ldots,\alpha_i$ satisfies F iff $\alpha_1^*,\ldots,\alpha_i^*$ satisfies F^*. Therefore, necessarily, $F\alpha_1,\ldots,\alpha_i$ iff $F^*\alpha_1^*,\ldots,\alpha_i^*$.[42] (Call this the *basic equivalence*.) Hence, the atomic proposition $[F\alpha_1,\ldots,\alpha_i]$ is necessarily equivalent to the atomic proposition $[F^*\alpha_1^*,\ldots,\alpha_i^*]$. Suppose that F is not the property of being manifest; i.e., suppose that $F \neq M$. Then, by definition of $*$, $*([F\alpha_1,\ldots,\alpha_i]) = [F^*\alpha_1^*,\ldots,\alpha_i^*]$. Thus $[F\alpha_1,\ldots,\alpha_i] \approx_N *([F\alpha_1,\ldots,\alpha_i])$. For our next case, consider any atomic proposition whose predicate is M and whose subject z is anything that is not a complex 1-ary intension. Suppose that z is a non-property or that $z = M$, then $[Mz]$ is an impure proposition. Therefore, by definition of $*$, $*([Mz]) = [M^*z^*]$. However, by the basic equivalence, we know that $[Mz] \approx_N [M^*z^*]$. So $[Mz] \approx_N *([Mz])$. Or suppose that z is a property G that is distinct from M. Then, $[Mz] = [M[Gx]_x]$. By definition of $*$, $*([M[Gx]_x]) = [(\exists x)G^*x]$. By the basic equivalence, we know that, necessarily, for all x and x^*, Gx iff G^*x^*. Using this, we will show that $[(\exists x)Gx] \approx_N [(\exists x)G^*x]$. Suppose $[(\exists x)Gx]$ holds. Then, for some x, $[Gx]$ holds, and, hence, so does $[G^*x^*]$. And, thus, $[(\exists x)G^*x]$ holds. For the other direction, suppose that $[(\exists x)G^*x]$ holds. Then, for some x, $[G^*x]$ holds. But given our theorem that, for all x, $x = *(*(x))$, we know that $x = *(y)$ where $y = *(x)$ and, therefore, that $x = y^*$. Consequently, $[G^*y^*]$ holds and, hence, so does $[Gy]$. And thus, $[(\exists x)Gx]$

holds. Since these relationships hold necessarily, it follows that, necessarily, $[(\exists x)Gx]$ holds iff $[(\exists x)G^*x]$ holds. Thus, $[(\exists x)Gx] \approx_N [(\exists x)G^*x]$. However, it is a logical truth that $[(\exists x)Gx] \approx_N [M[Gx]_x]$. Hence, $[M[Gx]_x] \approx_N [(\exists x)G^*x]$. But $*([M[Gx]_x]) = [(\exists x)G^*x]$. Therefore, $[M[Gx]_x] \approx_N *([M[Gx]_x])$. Summing up, suppose that p is an atomic proposition whose predicate is either a property other than M or a relation, or suppose that p is an atomic proposition whose predicate is M and whose subject is anything that is not a complex 1-ary intension. Then $p \approx_N *(p)$. Beginning with this conclusion for the above types of atomic propositions, we can argue by induction on the complexity of the remaining types of propositions that $p \approx_N *(p)$ for all p. There are four cases: (1) molecular propositions, (2) general propositions, (3) atomic propositions in which the predicate is M and the subject is a complex 1-ary intension, (4) atomic propositions in which the predicate is a complex i-ary intension, $i \geqslant 1$. The ideas used in the proofs are straightforward variants of those used in the proofs for the initial atomic cases.

It is easy to check that $*$ leaves all physical properties and relations unchanged, and it leaves a number of logical properties and relations unchanged, e.g., M, necessity, $=$, $\approx_N$, and $*$ itself. But can we be sure that $*$ alters the standard psychological relations; i.e., can we be sure that $R \neq *(R)$ for the standard psychological relations R?[43] (An affirmative answer is required for at least one such R if our construction is to refute functionalism.) To see that the answer is affirmative, consider the standard propositional attitude thinking. The generalized Quinean transformation T insures that $p \neq p^*$ for every pure proposition whose decomposition tree contains a formal node occupied by M or by existential generalization. Consider any such p. We know that, since $p \neq p^*$, it is not necessary that x thinks p iff x thinks p^*. At the same time, we know by the basic equivalence that, necessarily, x thinks p iff x thinks* p^*. It follows, therefore, that thinking and thinking* can have divergent extensions and, hence, that they are significantly distinct relations. Of course, thinking and thinking* can be multiply embedded within propositions that someone might think or think*. Therefore, the fact that thinking and thinking* are distinct produces an explosion in the possible ways the extensions of thinking and thinking* can diverge. And this, in turn, produces a chain reaction, namely, further possible ways in which the extensions of other propositional attitudes and of their $*$-counterparts can diverge. Furthermore, the distinction between thinking and thinking* guarantees that the extensions of introspecting and introspecting* can be distinct. The reason is this. A person x can introspect that xRy only if R is one of the conscious operations of mind (e.g., thinking); however, for suitable y, a person x can introspect* that xRy, where R is one of the $*$-counterparts of a conscious operation of mind (e.g., where R is thinking*). Likewise, the distinction between thinking and thinking* guarantees that, given the traditional empiricist theory of experience, the extensions of experiencing and experiencing* can be distinct. For although a person can experience the conscious operations of mind, he cannot experience their $*$-counterparts; however, a person can experience* the $*$-counterparts of the conscious operations of mind.

The operation $*$ has been constructed so that we can prove that the system of $*$-relations satisfies all the psychological principles we have been discussing in this paper, including even those for the self-conscious rational mind. To begin with, the proofs for the principles in the elementary belief/desire model are virtually identical to those given in the previous section except that (1) we make use of the fact that $p \approx_N *(p)$ where previously we made use of the fact that $p \approx_N t(p)$ and (2) we make use of the fact that xS^*p iff $xS*(p)$ where previously we made use of the fact that xS^*p iff $xSt(p)$. To see how the proofs go for the principles characterizing the self-conscious rational mind, let us start with the Infallibility principle. Given that principle, we wish to derive the following for arbitrary p:

If x thinks* that x thinks* p, then x thinks* p.

But, given our lemma and given that $p^* = q$, for some q, we see that this is actually equivalent to the following instance of the original Infallibility principle:

If x thinks that x thinks q, then x thinks q.

Thus, if the latter conditional is true, so is the former. And this completes the proof. Next consider the Introspection principle. Given this principle, we wish to derive the following for arbitrary p:

If x is thinking* p and x is entertaining* the proposition that x is thinking* p, then x introspects* that x is thinking* p.

However, given our lemma, and given that $p^* = q$, for some q, this is just equivalent to the following instance of the original Introspection principle:

If x is thinking q^* and x is entertaining the proposition that x is thinking q^*, then x introspects that x is thinking q^*.

Hence, if the latter conditional is true, so is the former. For a final example, consider the a priori Knowledge principle. Given this principle, we want to derive the following for arbitrary p:

If p is necessary and x clearly and distinctly understands* p, then x knows* p.

But given our lemma and given that $p^* = q$, for some q, this is just equivalent to the following instance of the original a priori Knowledge principle:

If q is necessary and x clearly and distinctly understands q, then x knows q.

Thus, if the latter conditional is true, the former is too. Furthermore, these results can be generalized in a a variety of ways. Indeed, our situation is quite analogous to the one we were in when we were generalizing our results concerning the elementary belief/desire model. In particular, we may use each of the resources developed in that context in order to obtain generalizations of comparable strengt' . (We might need to adjust the details to handle certain nonstandard logical or philosophical theories.) The conclusion is that the principles of psychology–even when they include principles for a so-

phisticated self-conscious rational mind–do not implicitly define the standard mental relations.

Do the *-counterparts of the standard psychological relations have the same causal roles as the standard relations? By an argument like that used in the previous section, we can show that every event of thinking* is an event of thinking and every event of thinking is an event of thinking*. Likewise for every other standard psychological relation and its *-counterpart. Thus, there exists no asymmetry in causal roles.[44] We have a complete system of relations that is functionally indistinguishable from the system of standard psychological relations.

THE STATUS OF MIND

We have a complete system of inverted relations that satisfies all the psychological principles satisfied by the standard psychological relations. It follows, therefore, that the principles of psychology do not implicitly define the standard psychological relations and, hence, do not provide a basis for direct functional definitions of them either. Moreover, because of their special modal ties to the standard psychological relations, these inverted relations cannot be distinguished from the standard relations in terms of causal role. Causally, the inverted relations are completely indistinguishable from the standard relations. Thus, the primary thesis of functionalism is untenable.

Those who are attracted to functionalism have two choices at this juncture. They can *abandon* functionalism and begin anew their search for a satisfactory theory of mind. Or they can *revise* their theory within a more powerful logical and metaphysical framework, namely, one with the ontology of qualities and connections. I will now say a few words on how this revision would go and on the metaphysical picture that results.

Consider the difference between green and blue, on the one hand, and grue and bleen, on the other. We can give logically necessary and sufficient conditions for grue and bleen in terms of green and blue, and we can give logically necessary and sufficient conditions for green and blue in terms of grue and bleen. However, only green and blue are genuine *qualities*. Grue and bleen are not; they are mere "Cambridge properties." Therefore, green and blue are logically and metaphysically prior to grue and bleen. Accordingly, on a logically and metaphysically strict conception of definition, grue and bleen are definable in terms of green and blue, but green and blue are not definable in terms of grue and bleen. In the domain of relations there is an analogous distinction. Many relations are grue-like "Cambridge relations." Others, however, are logically and metaphysically basic. Only these are genuine *connections*. If we are limning the true and ultimate structure of reality, the canonical scheme is that of qualities and connections; Cambridge properties and relations are wholly derivative. Or so this theory goes.

We have stated logically necessary and sufficient conditions for thinking* in terms of thinking, and we have stated logically necessary and sufficient conditions for thinking in terms of thinking*. And likewise for the other standard psychological relations and their inverted counterparts. However, it does not follow that on a metaphysically and logically strict conception of definition the standard psychological relations are definable in terms of their inverted counterparts. For just as green and blue are genuine qualities (namely, sensible qualities) and grue and bleen mere Cambridge properties, so the standard psychological relations might be genuine connections and their inverted counterparts mere Cambridge relations. Let us suppose that this is so. Then, by revising the earlier functionalist definitions, one can avoid the threat posed by the inverted relations. The revision consists of adding to the earlier definitions a new clause requiring that all the key relations involved be genuine connections.

This revison has two philosophically significant consequences, however. First, it would commit its proponents to the prospect of a purely logical analysis of intentionality and mind of the sort I have advocated elsewhere.[45] For that analysis begins with the premise that the standard psychological relations–or at least the basic psychological relations in terms of which the others are definable–are genuine connections and, hence, that they are logically and metaphysically basic. Then the analysis attributes to psychological connections certain characteristic logical properties, all of which are entailed by any moderately rich list of psychological principles. Thus, the revised definitions would incorporate both components of this purely logical analysis of intentionality and mind. Second, the revised definitions would rule out the prospect of saving any philosophically interesting form of materialism. (This prospect is what attracted many of the adherents of functionalism in the first place.) The reason philosophically interesting forms of materialism would be ruled out is that, given the revised definitions, the standard psychological relations would form a new category of connections that is logically different from the category of physical connections. For according to those definitions, the standard psychological relations would be identical to (or necessarily included in) connections that can contingently connect an individual to an intension independently of how that intension is realized in the world; as a categorial fact, physical connections can never have this distinctive logical character.[46] Therefore, since mental activity always involves mental connections, there is something essentially non-physical in all mental activity. But this is something that all philosophically interesting forms of materialism deny. In summary, perhaps the revised definitions succeed. But they entail an expanded conception of logic inasmuch as they commit one to the prospect of a purely logical analysis of intentionality and mind. And they entail an elevated, anti-materialist conception of the mind's metaphysical position in the universe.

More refined alterations of functionalism might be possible within this logical framework, but in the end the conclusion of the dialectic is always the same–no revised doctrine can succeed without elevating the mind well beyond any interesting materialist conception.

In informal terms what the existence of mind and anti-mind shows is this. We are in principle unable to identify the standard mental relations from the "outside,"

except perhaps by invoking a formulation that attributes to the mind a metaphysically basic, non-material character. In spite of this, we are certain that we can already identify these relations even in the absence of such a formulation. The only explanation is that we have access to the mind from the "inside."

Notes

1. George Myro, a brilliant philosopher and generous friend, has helped during each phase of my thinking about this topic. Donald Davidson, Jaegwon Kim, and Brian Loar made valuable points in preliminary conversations. Mark Bedau and David Reeve have contributed important comments and revisions, and Charles Wasson has done a dedicated and expert job editing, typing, and proofreading.

2. To illustrate, let us suppose that $A(S_1,\ldots,S_n)$ is the conjunction of the psychological principles that are supposed to define implicitly the standard mental relations $S_1,\ldots,S_n$. Let $A(R_1,\ldots,R_n)$ be the result of substituting variables 'R_1',..., 'R_n' for the constants 'S_1',..., 'S_n'. Then, the following would be a direct functional definition of the relation S_1 based on the theory A: $S_1 =_{df}$ the unique relation R_1 such that there exist unique relations $R_2,\ldots,R_n$ that, together with R_1, make the theory $A(R_1,\ldots,R_n)$ come out true when we hold constant the interpretation of the physical and logical constants in A. See my paper "An Inconsistency in Functionalism," *Synthese* 38 (1978): 333-72, for a discussion of the various kinds of direct functional definitions.

A word is in order about the identity conditions of properties, relations, and propositions. In this paper I will adopt the standard practice of treating properties as identical if they are necessarily equivalent and of treating relations (of equal degree) as identical if they are necessarily equivalent. However, I will treat propositions as more "fine-grained" than this, for in thought we can cut distinctions more finely than necessary equivalence.

3. Since such functional definitions would contain only physical and logical constants, functionalism is consistent with physicalism of the terminological variety.

4. I should add that the construction can be adapted to refute functional theories and reductionistic theories in any subject area where "fine-grained" intensionality plays a role. Thus, it can be adapted to refute many forms of functionalism and reductionism in linguistics, social and political theory, legal theory, ethics and aesthetics. And it can be adapted to refute many forms of the currently popular "causal theory" of reference, meaning, and mind.

5. There are some damaging technical criticims based upon the results given in my paper "An Inconsistency in Functionalism" and upon those given in S. Thomas, *The Formal Mechanics of Mind* (Ithaca, N.Y., 1978). I will not discuss these criticisms in this paper since they presuppose knowledge of recursive functions. However, I do wish to note that some commentators have failed to grasp the significance of these criticisms, perhaps as a result of their liberal use of mathematical logic. For example, Ned Block ("Introduction: What Is Functionalism?" *Readings in Philosophy of Psychology*, vol. 1 [Cambridge, Mass., 1980], 171-84) erroneously reports that the aim of "An Inconsistency in Functionalism" is to show that functionalism can be made to fit a certain formal definition of behaviorism, and then, without giving any argument, Block concludes that the technical result in the paper is "misguided" because it "blurs the distinctions between functionalism and behaviorism." This is seriously confused. First, my criticism of functionalism is expressly aimed at physiologically oriented functionalism as well as behaviorally oriented functionalism. Second, the aim of the paper is not at all to get functionalism to "fit" a definition of behaviorism; that would be silly. Nor are the distinctions between functionalism and behaviorism blurred. The technical result in the paper is a proof that a functional definition is *adequate* if and only if there also exists an adequate non-functional definition. Now functionalism is based upon a few well-known negative arguments to the effect that the non-functional definitions envisaged by behaviorists and psycho-physiological reductionists must always be inadequate. However, if these arguments are sound, we can easily generalize them so that they will apply to all the non-functional definitions of the kind dealt with in my technical result. (From a formal point of view, it makes absolutely no difference to the success of these negative arguments

that some of these non-functional definitions might be wholly impractical, and it makes no difference that some of these definitions might be more complex than those envisaged by many behaviorists.) Therefore, it follows that, if the functionalist's negative arguments are sound, then his proposed functional definitions can never be adequate. Hence, the inconsistency in functionalism. For some reason Block thinks that this criticism turns on what historically counted as behaviorism; but plainly that question is irrelevant.

6. For example, J. Paul Churchland, "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy* 77 (1981):67-90. For a powerful argument against eliminative materialism, see George Myro, "Aspects of Acceptability," *Pacific Philosophical Quarterly* 62 (1981):107-22.

7. For example, Tyler Burge, "Individualism and the Mental," *Midwest Studies in Philosophy* 4 (1979):73-122.

8. Such a theory is developed in sec. 39, "Pragmatics," in my book *Quality and Concept* (Oxford, 1982).

9. Ned Block, "Troubles With Functionalism," in *Perception and Cognition: Issues in the Foundations of Psychology*, edited by C. W. Savage, Minnesota Studies in the Philosophy of Science, vol. 9 (Minneapolis, 1978), 261-325.

10. Douglas R. Hofstadter, "Who Am I Anyway?" *New York Review of Books*, May 29, 1980.

11. See, for example, Ned Block and Jerry Fodor, "What Mental States Are Not," *Philosophical Review* 81 (1972):159-81; Sydney Shoemaker, "The Inverted Spectrum," *Journal of Philosophy* (1982):357-82.

12. There is some confusion over what the functionalist's thesis implies. It does not imply that sensible qualities (e.g., the various shades of red) have functional definitions; nor does it imply that such properties as the property of sensing-red or the property of having-an-experience-of-red are functionally definable. Rather it implies that, given an arbitrary quality F (e.g., a shade of color), the property of sensing F or the property of having an experience of F has a functional definition. This, in turn, implies that the relation of sensing or the relation of experiencing has a functional definition. We can, of course, experience "reflective" qualities (e.g., sadness), which are psychological by nature. The functionalist is committed to the functional definability of these special qualities. But can distinct reflective qualities be perfectly equivalent functionally? Perhaps, but the problem is that no one seems able to give a conclusive argument for this view, and there are some plausible considerations that seem to count against it. See, e.g., David Lewis, "Mad Pain and Martian Pain," in Block, *Readings*, 216-22.

13. Another candidate counterexample, which bears some resemblance to the inverted-spectrum example, is based on the alleged possibility of "absent qualia," i.e., the possibility that there could be a psychological state that is functionally equivalent to one involving, say, sensing red but that does not involve any sensible quality at all. (See Block and Fodor, "What Mental States Are Not"; Sydney Shoemaker, "Functionalism and Qualia," *Philosophical Studies* 22 [1975]:291-315, and "Are Absent Qualia Impossible–A Reply to Block," *Philosophical Review* 90 [1981]: 581-99; Ned Block, "Are Absent Qualia Impossible?" *Philosophical Review* 89 [1980]:257-74.) However, considerations analogous to those just given in connection with the inverted-spectrum example show that this type of counterexample is also inconclusive.

14. John R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3 (1980):417-24. I have heard of a somewhat related example, namely, that of some computer game software with the following intriguing property. When this software is used with one type of computer hardware, a certain game is "played"; but when it is used with another type of computer hardware, a quite distinct game is "played." Although this example is also very suggestive, it cannot be used to refute functionalism, for it is open to the same kind of rejoinder the functionalist may use against the Chinese-room example. In particular, following a program is not sufficient for *playing* a game. Playing a game requires *knowing* how to play the game and *intending* to do so. Using a chain of reasoning analogous to that used in reply to the Chinese-room example, we are driven to the following conclusion. To use the example in a refutation of functionalism, we must already have established the existence of a global system of new relations–knowing*, intending*, etc.–that is functionally isomorphic to, yet distinct from, the system of standard psychological relations–knowing, intending, etc. But if we could do this, we would already have succeeded in refuting functionalism quite

independently of the present example. So once again we are no closer to a refutation of functionalism.

15. For example, the versions of Lewis, Harman, Shoemaker, Loar, and others. It should also be emphasized that, just as functionalism is not tied to the Turing-machine doctrine, it is not tied to the language-of-thought thesis, i.e., the thesis that the immediate objects of thought are mere syntactic entities. David Lewis, for example, is adamantly opposed to this thesis. Incidentally, if the Chinese-room example were to refute Turing-machine functionalism, would it follow that we intentional beings are not Turing machines? Yes and no. The example would show that being a Turing machine is not a sufficient condition for intentionality, but it would do nothing to show that being a Turing machine is not a necessary condition. Thus, as far as the example is concerned, every intentional being might have to be a species of Turing machine, i.e., a Turing machine with some special additional feature.

16. Two observations are in order. First, in this paper I use 'thinks' to single out a central use of 'believes', viz., the use often conveyed in philosophical discussion by 'occurrently believes'; I do not use 'thinks' in the sense of 'entertains'. Second, in the definitions of the inverted relations, '$t(p)$' is intended to have narrow scope.

17. In symbols, $(p \;\&\; \boxed{C}\,(p \leftrightarrow x \text{ entertains } p)) \rightarrow x$ thinks p. This is a coarse approximation of a causal theory of perceptual belief.

18. For example, we could give the construction in the setting of a "world-relative" and time-relative treatment of identity: accordingly, relative to any given "possible world" and any given time, $h_x(v)$ would be identified with an appropriate ordinary object or region of space. Or we could rephrase the second clause in the definition of h_x thus: $h_x(v) =_{df}$ the object u such that, if x exists, u is the ordinary object (if there is one; otherwise, u is the region) that is located at the same distance d ($d \geqslant 0$) from the plane $l(x)$ as v but on the opposite side, and if x does not exist, $u = v$. Then in the definition of t_x we would treat the notation 'a_i'', $i \geqslant 1$, as a defined descriptive expression: $a_i' =_{df} h_x(a_i)$.

19. Any question about the uniqueness of $h_x(v)$ can be resolved by any of a number of standard techniques. For example, if v is an ordinary physical object, let $h_x(v)$ be the mereological union of the objects satisfying the remainder of the defining condition for $h_x(v)$; and if $v = h_x(v')$ for some ordinary physical object v', let $h_x(v) = v'$.

20. That is, the following principle is valid for all logical operators Θ in T: if it is necessary that A iff B, then $\Theta(A)$ iff $\Theta(B)$.

21. This complementarity neutralizes the functional significance of differences in logical form between, say, particularese propositions and universalese propositions.

22. The following is a principle of this type that fails to be satisfied by thinking*: if x is a *Homo sapiens* (or if x is George Bealer) and x's eyes are focused on x's moving left hand, then, in all probability, x will think that it moves.

23. Another functionalist response would be to restrict the range of the relations over which he quantifies in his candidate functional definitions; specifically, he might restrict the eligible relations to those that are "internal." (See, e.g., Ned Block, "Psychologism and Behaviorism," *Philosophical Review* 90 [1981]:5-43.) However, this response is of no help if '**internal**' is intended to mean either physically internal or causally independent of things "outside" (e.g., physically separate from) the individual agent or anything in this vein. For our inverted relations thinking*, etc. are in these respects every bit as internal as the standard propositional attitudes. Like th standard propositional attitudes, the inverted relations are realized "in" us all the time. Indeed, it is necessary that xSp iff $xS^*t(p)$ and $xSt(p)$ iff xS^*p, for every particular x, every proposition p, and every standard propositional attitude S and its inverted counterpart S^*. So, for example, every time you think that there exists a rabbit, you are also thinking* that rabbithood is manifest. And this is so whether or not you realize it. On the other hand, if 'internal' is intended to mean "psychologically internal" in the sense of being present to the mind, the present response only undermines functionalism. For the central thesis of functionalism is that the standard psychological relations are functionally definable without recourse to psychological terms. But 'internal' in the last sense is patently psychological.

24. If t is the particular/universal transformation, something even stronger holds. The following will help to show what it is: for any sentence A, there is a sentence B such that it is provable that, necessarily, the proposition that $B = t(\text{that } A)$ and, in turn, it is provable that, necessarily, x thinks that B if and only if x thinks $t(\text{that } A)$.

25. Some more complex examples will help to dramatize the point. Thus, t(the proposition that the property of hitting someone is manifest) = the proposition that someone is such that the property of being hit by him is manifest. In symbols, $t([M[(\exists y)Hxy]_x]) = [(\exists x)M[Hxy]_y]$. Or t(the proposition that someone hits someone) = the proposition that the property of being someone such that the property of being hit by him is manifest is itself a manifest property. In symbols, $t([(\exists x)(\exists y)Hxy]) = [M[M[Hxy]_y]_x]$. Or just try the proposition that someone gives something to someone; in symbols, $t([(\exists x)(\exists y)(\exists z)Gxyz]) = [M[M[M[Gxyz]_z]_y]_x]$. Clearly, we can think such particularese propositions without even being able to understand their universalese counterparts. Of course, with practice we could readily think the latter, and no doubt there could be beings who naturally think the latter but have trouble thinking the former.

26. This point was inspired by conversation with Donald Davidson.

27. The following example makes the same point even more directly. Suppose that type-a beings think S iff physical inscriptions of the pure abstract shape S occur in the "thinking centers" in their brains and that they think $t(S)$ iff physical inscriptions of the pure abstract shape $t(S)$ occur there. And suppose that type-b beings think S iff physical inscriptions of the pure abstract shape $t(S)$ occur in their "thinking centers" and that they think $t(S)$ iff physical inscriptions of the pure abstract shape S occur there. According to functionalism, since S and $t(S)$ have a direct structural isomorphism to one another (i.e., the function t itself), the causal role of inscriptions of S in type-a beings would be identical to the causal role of inscriptions of $t(S)$ in type-b beings, and the causal role of inscriptions of $t(S)$ in type-a beings would be identical to the causal role of inscriptions of S in type-b beings. Therefore, according to functionalism, type-a and type-b beings would be functionally indistinguishable. Thus, a physical inscription in a being's "thinking center" of a Mentalese sentence S (i.e., a physical inscription of the pure abstract shape S) is on its own no indication of whether the being is thinking S or whether it is thinking $t(S)$ instead. (Of course, it is far-fetched to think that physical inscriptions of the pure abstract shapes S and $t(S)$ would occur in any more than the smallest fraction of the possible beings who think S or think $t(S)$.)

28. Of course, the fact that actual physical inscriptions of syntactic entities can occur in the brains of these other types of beings is of no help. For just as type-a and type-b form a functionally indistinguishable complementary pair, these other types also come in functionally indistinguishable complementary pairs. For example, type-c beings might think S iff an inscription of the sentence B in langauge L occurs in their "thinking centers," and they might think $t(S)$ iff an inscription of $k(B)$ occurs there. And type-d beings might think S iff an inscription of $k(B)$ occurs in their "thinking centers," and they might think $t(S)$ iff an inscription of B occurs there. (This function k is defined on L in a way analogous to the way g was defined on English.)

29. Incidentally, we are now in a position to identify an inherent defect in the language-of-thought theory itself. I can think that A, that B, that C, and so on. It is certainly possible to hypothesize ideal languages such as Mentalese, and we can easily define some mapping r from the English expressions 'that A', 'that B', 'that C', etc. onto sentences in Mentalese. However, we can just as easily define an alternative mapping s such that $s(x) = t(r(x))$ always holds. Now with which sentence in Mentalese should we identify the entity denoted by 'that A'? Should it be r('that A') or s('that A')? The choice is utterly arbitrary. This arbitrariness is damaging just in its own right. However, I believe that it becomes absolutely fatal when language-of-thought functionalism tries to define the relationships between thought and sensation and between thought and the properties of external objects. To succeed at this, one must reject the representationalism implicit in the language-of-thought theory and adopt instead a form of realism.

30. There is some question about whether it is valid to expand the embedded occurrences of 'thinking*'; for example, such expansion overlooks the phenomenon of the paradox of analysis. However, a moment's reflection will show that our original construction is beset with the same problem whether or not we expand these occurrences of 'thinking*'. So for simplicity of exposition I will do so in the text.

31. On one treatment of functional constants, 'f' is contextually defined as follows (where 'F' is an appropriate 2-place predicate): $x < f(y)$ iff$_{df}$ $(\exists_1 z)\,(Fyz \;\&\; x < z)$. On this treatment, therefore, the proposi-

tion that $x <f(y)$ would be an existential proposition. Thus, if we were to adopt this treatment, my argument in the text would need to be complicated somewhat. But the main point would remain the same: t(that $x <f(y)$) would not contain z as a constituent even though $f(y) = z$; on the contrary, it would continue to contain y as a constituent, just as it did prior to being transformed by t. And this is where the problem lies with all the "superficial" Quinean transformations: the embedded propositions remain unchanged.

32. In a comprehensive formulation, the above principles of Introspection should be supplemented with the following companion principle: for any relation R, if x introspects that xRy, then xRy.

33. We should try to construct a new relation experiencing* such that it is possible to experience* such things as thinking*, desiring*, deciding*, etc. but not thinking, desiring, deciding, etc. However, we must do this in such a way that it would be possible to experience* experiencing* itself but not to experience* experiencing. It is a challenge to construct a new relation with this special kind of impredicativity.

34. We should also take into account some version of the following companion principle:

> If p is not contingent and x clearly and distinctly understands p and x believes p, then p is necessary.

35. For example:

> If x stands in the relation of clear and distinct understanding to the proposition that knowing $\neq a$, then x stands in the relation of knowing to the proposition knowing $\neq a$

where a rigidly designates knowing*. To test whether the inverted relations of clear and distinct understanding* and knowing* satisfy this principle, substitute 'clear and distinct understanding*' for 'clear and distinct understanding' and 'knowing*' for 'knowing'. We obtain:

> If x stands in the relation of clear and distinct understanding* to the proposition that knowing* $\neq a$, then x stands in the relation of knowing* to the proposition that knowing* $\neq a$.

(Note that we do not substitute anything for the rigid designator 'a'. The reason is that 'a' refers directly to the relation of knowing* quite independently of the fact that we chose to define knowing* in terms of knowing. For in the present context we are proceeding under the standard assumption that relations are identical if necessarily equivalent and, hence, that each relation has an infinite number of necessarily equivalent definitions, none more basic than another.) Given our definitions of the inverted relations, however, we know that this is equivalent to the following:

> If x stands in the relation of clear and distinct understanding to the proposition that knowing* $\neq a$, then x stands in the relation of knowing to the proposition that knowing* $\neq a$.

But this outcome is plainly wrong. No one can know something that is false, and it is certainly false that knowing* $\neq$ a. Thus, we see that, when the inverted relations knowing* and clear and distinct understanding* take the place of knowing and clear and distinct understanding, respectively, they do not satisfy the original principle. It will occur to the reader that we might save our construction from this problem as follows. First, define a third relation knowing**. Then, show that when the three relations clear and distinct understanding*, knowing*, and knowing** take the place of clear and distinct understanding, knowing, and knowing*, respectively, they do succeed in satisfying the original principle. However, when generalized, this strategy forces us to posit an infinite hierarchy of mutually dependent relations: knowing, knowing*, knowing**, knowing***,...; believing, believing*, believing**, believing***,...; and so on. And it turns out to be extremely difficult to construct such a hierarchy. Furthermore, if we are not careful, the identity of the standard psychological relations might be uniquely determined by their "functional roles" (i.e., their ground-level positions) in this hierarchy, thus validating a version of functionalism. These, then, are some further problems a new construction must overcome.

36. See sec. 42, "Realism and Representationalism," *Quality and Concept*, for more on the realism/representationalism controversy.

It takes a special logical framework to treat non-elementary psychological principles. This logic must

be equipped to treat liberal forms of quantifying-in. It must cut the "fine-grained" intensional distinctions characteristic of intentional matters, and at the same time it must have an apparatus for treating necessary truths. It must be able to represent relations (e.g., thinking) whose objects are propositions that can "contain" those very same relations as "constituents," and it must be able to represent relations (e.g., identity, thinking of, experiencing, etc.) that can fall within their very own ranges. Finally, it must be able to represent "transcendental" relations, i.e., relations (such as identity, thinking of, introspecting, experiencing) whose ranges span more than one ontological category. As far as I know, *Quality and Concept* provides the only philosophically satisfactory logical framework that has all these special features. So I will take the liberty to allude to it in what follows.

37. For example, consider a psychological theory consisting of the elementary belief/desire model plus all the principles for the self-conscious rational mind listed above. Let this theory be formulated in the intensional logic T2 (see sec. 16, *Quality and Concept*) supplemented with predicates for primitive physical and mathematical relations $O_1,\ldots, O_m$ and primitive psychological relations $S_1,\ldots, S_n$. Let ‹M, I› be a model for the resulting theory P in which the physical and mathematical predicates are given their intended interpretation. To show that P does not implicitly define $S_1,\ldots, S_n$, we construct an alternate model structure M^* that has two systems of relations $S_1,\ldots, S_n$ and $S_1^*,\ldots, S_n^*$ but that is identical to M in its physical and mathematical components. To do this, adjoin to the domain D of M any new primitive relations $S_1^*,\ldots, S_n^*$, and then close the new domain D^* under the extended logical operations Conj*, Neg*, Exist*, etc. [E.g., Conj* is such that: (1) if $x, y \in D_i$ for $i \geq 0$, then $\text{Conj}^*(x,y) = \text{Conj}(x,y)$ and (2) if $x, y \in D_i^*$, for $i \geq 0$, and $x \notin$ D or $y \notin$ D, then a new element z is added to D_i^* and $\text{Conj}^*(x,y) = z$.] Next define the alternate extension functions $H^* \in K^*$ in M^* in terms of the alternate extension functions $H \in K$ in M: (1) $H^*(x) = H(x)$ if x is either an individual or $O_1,\ldots, O_m$ or $S_1,\ldots, S_n$; (2) $H^*(\text{Id}) = \{uv \in D^*: u = v\}$; (3) $H^*(S_i^*) = \{uv: \langle u, *(v)\rangle \in H(S_i)\}$, where $*(v) =_{df}$ the item w in D^* that is just like v except that, for every occurrence of S_i, $1 \leq i \leq n$, in v's decomposition tree, there is an associated occurrence of S_i^* in w's decomposition tree. For each of the remaining items $x \in D^*$, $H^*(x)$ is determined by the standard conditions characterizing the behavior of the fundamental logical operations. Finally, the actual extension function G* in M^* is the function H^* in K^* that is associated by the above definition with the actual extension function G in M. With M^* so constructed, it is easy to prove a lemma that, for every proposition p in D^*, p and $*(p)$ are necessarily equivalent. Now let I^* be an interpretation that is just like the original interpretation I except that I^* assigns S_i^* to a predicate in P iff I assigns S_i to that predicate. Given the above lemma, it is then simple to show that $\langle M^*, I\rangle$ and $\langle M^*, I^*\rangle$ both satisfy P. Hence, P does not implicitly define the standard psychological relations $S_1,\ldots, S_n$ even when the interpretation of its physical and mathematical predicates is held constant. And much stronger results of this general type are also possible.

38. That is, a framework of intensional logic that has not yet been characterized axiomatically. It is well known that in intensional logic, just as in set theory, the use of certain pathological principles leads to paradoxes. No one has yet found a systematic method for weeding out only pathological principles so that the sound ones can be set forth axiomatically. But this does not invalidate arguments that employ sound principles. Recall, moreover, that the reason we find ourselves in this territory is that we are trying to accommodate *the functionalist's* request for a logical framework that can handle the special forms of impredicativity present in his various non-elementary psychological principles.

39. A similair example is the property of being someone who is thinking of the thinking-of relation. This property is the result of predicating the thinking-of relation of itself; symbolically,

$$[\text{u thinks of } [\text{u thinks of v}]_{uv}]_u = \text{Pred}([\text{u thinks of v}]_{uv}, [\text{u thinks of v}]_{uv}).$$

Another example involving self-predication is a psychological property encountered earlier, namely, the property of being someone who experiences experiencing itself. This property is the result of predicating the experiencing relation of the experiencing relation itself; symbolically,

$$[\text{u experiences } [\text{u experiences v}]_{uv}]_u = \text{Pred}([\text{u experiences v}]_{uv}, [\text{u experiences v}]_{uv}).$$

40. The notation for unordered pairs is defined thus:

$$[x,y] =_{df} [v = x \text{ or } v = y]_v.$$

Then the notation for finite sequences is defined as follows:

$$\langle\alpha_1\rangle =_{df} \alpha_1;\ \langle\alpha_1, \alpha_2\rangle =_{df} [[\alpha_1, \alpha_1], [\alpha_1, \alpha_2]];\ \langle\alpha_1,\ldots, \alpha_{i+1}\rangle =_{df} \langle\langle\alpha_1,\ldots, \alpha_i\rangle, \alpha_{i+1}\rangle.$$

For simplicity I often omit the use of the angle brackets in the text.

41. For an explanation of why this is so and for a general defense of first-order constructions, see chap. 5, "Predication," *Quality and Concept*.

42. This step in the proof relies on the intuitive assumption that, for each F and F^*, the following two abstraction principles either are both valid or are both not valid: (a) $F\alpha_1,\ldots, \alpha_i$ iff $\alpha_1,\ldots, \alpha_i$ satisfies F, and (b) $F^*\alpha_1^*,\ldots, \alpha_i^*$ iff $\alpha_1^*,\ldots, \alpha_i^*$ satisfies F^*. Certain instances of these abstraction principles give rise to a familiar problem: for pathological relations F and F^*, (a) and (b) are inconsistent with principles of classical logic. Although we still await an ideal solution to these paradoxes, this does not detract from the soundness of my argument. For my argument assumes only that the validity of (a) and the validity of (b) stand or fall together. A survey of examples indicates that F and F^* are both non-pathological or they are both pathological. If the former, then (a) and (b) would both be unproblematically valid. If the latter, then special steps must be taken. One approach would be to deny the relevant instances of (a) and (b); another approach would be to depart from principles of classical logic, e.g., by permitting truth-value gaps for the relevant pathological propositions. However, whichever approach is taken, symmetry demands that it be taken for both (a) and (b). Thus, whether or not F and F^* are pathological, (a) and (b) would stand or fall together; i.e., they would both be valid or they would both fail to be valid. Therefore, we may expect that this symmetry will be preserved in any ideal solution to the paradoxes. Indeed, when we speak of an ideal solution, we mean one in which symmetries of this sort are preserved. Therefore, even though we do not yet possess an ideal solution, we have good reason to expect that it will validate the assumption used in our proof.

Of course, the whole issue might be avoided simply by weakening our lemma to propositions not dependent on the pathological cases. Since it is implausible that the truth of functionalism could turn on psychological "laws" for how we think about pathological cases, such a weakened lemma should suffice in the remainder of our argument.

43. The sensing relation is one exception, for we have not built into the definition of * a function f that "inverts" some spectrum of sensible qualities. This could be done, but it would invite special controversies that we should be wise to avoid here. After all, our construction already refutes functionalism. Moreover, given the traditional empiricist theory of experience, sensing is only a mode of experiencing, and experiencing has been shown to be quite distinct from its inverted counterpart experiencing*.

44. Are the *-counterparts of the standard psychological relations "internal" relations? By an argument just like the one used in note 23, we can show that, in any sense of "internal" available to the functionalist, these *-relations are internal. Necessarily, you are thinking* p^* whenever you are thinking p, and this is so whether or not you realize it!

45. Chap. 10, "Mind," *Quality and Concept*.

46. These claims are examined more thoroughly, ibid.

MIDWEST STUDIES IN PHILOSOPHY, IX (1984)

Parallelism, Interactionism, and Causation

LAIRD ADDIS

One may gather from the arguments of two of the last papers[1] published before his death that J. L. Mackie held the following three theses concerning the mind/body problem:

(1) There is a distinct realm of mental properties, so a dualism of properties at least is true and materialism false.
(2) All bodily movements probably have sufficient causes in physical facts and properties, but mental facts and properties are not causally irrelevant to human action.
(3) At the same time, the view that there are not sufficient causes in the physical realm alone for all bodily movements has no good and adequate empirical or philosophical reasons against it.

In this paper I wish (1) to register my strong agreement with the first thesis by way of simply taking it for granted, (2) to defend the second thesis in greater detail and in a manner somewhat different from Mackie's, and (3) to show the third thesis to be false.

I

If a dualism of properties is true, there are fundamentally three abstract possibilities: the mental properties are related to the crucial physical properties by (1) laws of coexistence, (2) laws of succession, or (3) no laws at all. These views may reasonably be labeled as (1) *parallelism*, (2) *interactionism*, and (3) *fatalism*, respectively. Although there have been people who, crippled by their theological commitments, have thought they believed in fatalism, it is phenomenological absurdity to maintain that what one desires or chooses or values never *makes a difference* to one's behavior; and no one ever acts that way either (whatever it could