J. A. FODOR

Computation and Reduction

Introduction

Nobody loses all the time; a fitful and negative glimmer illuminates the philosophy of mind. We now know that the program of behavioristic reduction of psychological theories cannot, in general, be carried through. And we know why. Let a behaviorist be someone who claims, at a minimum, that there must be a reference to behavior in any logically perspicuous specification of a psychological state. Then, whatever else may be wrong with the behaviorist program (no doubt plenty else is wrong), it is blocked by the intentionality of typical psychological terms.[1] So, for example, there is a *sense* in which there is a reference to whistling (to, as psychologists inelegantly say, Dixie-whistling behavior) in such English formulae as ⌜Ψ = John's intention to whistle a snatch of Dixie⌝, and perhaps there would be a similar sort of reference to, say, avoidance behavior in any adequately perspicuous specification of John's pains. Perhaps, that is, John's having a pain profoundly involves his intending or desiring to avoid a painful stimulus. But, of course, that would not be good enough to make the behaviorist's case. For such "references to behavior" as are, in *this* sense, involved in logically perspicuous specifications of psychological states occur (only) in intentional contexts; which is to say that they are not, in any full-blooded sense, references to behavior at all. Briefly, whatever is required to make it true that John's intention to whistle a snatch of *Dixie* is such and such does not in general, involve the actual occurrence of any *Dixie*-whistling, and whatever is required to make it true that John's pain is such and such does not, in general, involve the actual occurrence of any avoidance behavior. Similarly, mutatis mutandis, for other psychological states and processes.

This is all familiar territory, and I mention these points only in order to set them to one side. My present interest is the following. Although it is fairly clear what kinds of problems phenomena of intentionality raise for *behavioristic* reductions of psychological predicates, very little is understood about the problems they raise for *physicalistic* reductions of the sort often contemplated by central-state identity theorists. In fact, it often happens in the standard literature on mind/body identity that this question is not so much as aired.[2] Perhaps this is due to the continuing influence of an early version of the identity theory, which was physicalist about *sensations* but behaviorist about propositional attitudes (see, for example, Place, 1956 and Smart, 1957). On that view, physicalism presupposes behavioral analyses for those psychological predicates that most evidently establish intentional contexts: verbs like "hopes," "thinks," "intends," "feels that," "believes," etc. The identity theory is thus left free to operate in the account of sensations, an area where issues of intentionality seem less pressing.

But however one construes the history, it now seems clear that behavioristic analyses of propositional attitudes will not be forthcoming, so physicalist theories will have to decide what to do about their intentionality. And it also seems clear that the problems intentionality poses for physicalism are likely to be quite different from the ones it posed for behaviorism. The reason for this is that, prima facie, what the behaviorist requires of reduction is quite different from what the physicalist requires. Behaviorism will not do unless for every true sentence in an (ideally completed) psychology there exists a canonical *paraphrase* in a proprietary vocabulary. Nonlogical expressions in this vocabulary are to be behavioral (whatever, precisely, that is supposed to mean), and all contexts in canonical paraphrases are to be extensional.

It is a good deal less clear what the physicalist wants. *Not* a paraphrase, surely, because on no plausible account does a physicalistic sentence say what the corresponding mentalistic sentence means. But, on the other hand, the physicalist presumably wants something more than an extensional sentence materially equivalent to each intentional sentence, since that is available by merest stipulation. Let 'Alfred = John's intention to whistle a snatch of *Dixie*' be a stipulated equivalence. Then, for 'John intends to whistle a s. of D.' we

get, roughly, 'Here's Alfred,' which satisfies none of the traditional tests for opacity. Nor will it content the physicalist to find an extensional sentence equivalent to each intentional one such that the equivalence is nomologically necessary since, presumably, a psychophysical parallelist or epiphenomenalist could grant him that without granting him what he primarily wants. Nor, I think, should the physicalist be content with the de facto identity of the things that mental terms name with those that physical terms name since, as we shall see, there are several respects in which he might get that and not get the substantive reduction of psychology to neurology (or any more basic science). In fact, it seems difficult to me to say just what the physicalistic reductionist *does* want. In this paper, in any event, I shall consider what, in the light of the intentionality of typical psychological predicates, he is likely to get, and what it is likely to cost him.

My strategy will be the following. I shall say a few things about reduction and a few things about psychological explanation. I shall then argue that, given the notions of reduction and psychological explanation at issue, the reduction of psychology could probably be purchased only at the expense of its explanatory power. I shall suggest, too, that this situation is probably specific to psychology as opposed to other special sciences. That is, reducing psychology to (say) neurology would probably lose us something that reducing meterology to (say) mechanics would probably not lose us. This argument will turn on the special role that intentional expressions play in psychological theories. Finally, I shall discuss very briefly how we could strengthen the notion of reduction so as to guarantee that, if psychology is reducible in this stronger sense, then it *is* reducible without loss of explanatory power.

I shall not, however, argue for or against the blanket contention that psychology is reducible to neurology or physics. Indeed, it is hard to see how a sensible argument to either conclusion *could* be mounted in the present state of play. For, not only are there straightforwardly empirical questions that are pertinent and unanswered, but also what we say about the reducibility of psychology depends on what we think psychological theories should be like and what we require reduction to preserve. And we do not know much about what psychological theories should be like or about which constraints on reduction are justifiable.

Reduction

In what follows, I shall be taking "reduction" in quite a specialized sense; the sense that (if I read the literature correctly) informed much positivistic thinking about the relation between physics and the special sciences. It will be one of my points that this notion of reduction is not the only one compatible with the ontological assumptions of physicalism. But I shall start with it because it is widely known and because the considerations that are likely to make psychological theories recalcitrant to reduction in this special sense would also hold on many other construals, as far as I can tell. So as not to have to write "reduction in this special sense" whenever I wish to refer to reduction in this special sense, I shall adopt the practice of calling it simply "reduction." But the reader is encouraged to bear the qualifier in mind.

In the first place, then, "reduces to" names a relation between *theories*. When that relation holds between a pair of theories, say T_1 and T_0, T_0 is said to be a *reducer* of T_1. The reduction relation is transitive and asymmetrical, hence irreflexive. By the "unity of science" I shall mean the doctrine that all sciences except physics reduce to physics. By "physicalistic reductionism" I shall mean a certain claim that is entailed by, but does not entail, the unity of science; viz., the claim that psychology reduces to physics (presumably via neurology).

I do not know whether theories are sets of sentences, and I do not wish to prejudice that question. However, some of the conditions on reduction constrain properties of the syntax and vocabulary of expressions in the reduced and reducing sciences. So, in what follows, I shall take "theory" to mean "theory in normal form." A theory in normal form *is* a set of sentences by stipulation. I shall also sometimes write as though all the sentences that belong to a theory in normal form are universal generalizations. The interesting problems about reduction persist on this assumption, and it helps with the exposition.

Let each sentence of the set T_1 be the universal closure of a formula of the form $A_x \rightarrow B_y$ (read: it is a law that x's being A is causally sufficient for y's being B). Let each sentence of the set T_0 be the universal closure of a formula of the form $\Phi_x \rightarrow \Psi_y$. Then the crucial conditions on T_0 reducing T_1 are these.

a) (At least some) items in the vocabulary $A, B \ldots$ are not in the vocabulary $\Phi, \Psi \ldots$

b) Let the "projected" predicates of a science be the ones that appear essentially in its laws. Then T_0 reduces T_1 only if nomologically necessary and sufficient conditions for the satisfaction of the projected predicates of T_1 can be framed in the vocabulary of T_0.

So, for example, theories in whose laws the expression 'water' (or its cognates) occurs will reduce to chemistry only if (a′) chemistry contains some expression other than 'water' (say 'H_2O') such that (b′) '(x) (x is water iff x is H_2O)' expresses a law.

c) Let T_2 be the set whose members are T_0 together with such laws. Then T_0 reduces T_1 only if every consequence of T_1 is a consequence of T_2.[3]

Formulae like the one quoted in (b′) are said to express "bridge" laws; we can call them *bridge formulae*. For our purposes, the essence of standard reductionism is the suggestion that bridge formulae link reduced sciences to their reducers. Viewed as principles of inference, bridge formulae permit us to substitute expressions in the vocabulary of T_0 for expressions in the vocabulary of T_1 preserving nomological necessity. That is, if G is a generalization in the vocabulary of T_1, and G' is the formula derived from G by replacing every expression of T_1 by the T_0 expression that is related to it by some bridge formula, then if G is nomologically necessary, G' will be too. For essentially this reason, it is plausible to claim that any event causally explained by G is also causally explained by G'.

There are many difficulties with this notion of reduction, but I shall not pursue them here. Suffice it to mention only the following rather general points. First, on this account, reducibility involves a good deal more than the ontological claim that things that satisfy descriptions in the vocabulary of T_1 also satisfy descriptions in the vocabulary of T_0 (e.g., that every event that falls under a law of psychology satisfies a physical description). I have stressed this point elsewhere, but it may be worth repeating here in passing. Since the present account requires that bridge formulae be lawlike, it entails not only that their antecedents and consequents be expressions of the reduced and reducing science respectively, but also that they be *projectible* expressions of the reduced and reducing sciences respectively. (This is of a piece with the remark made above, that

substitution under the equivalences specified by bridge formulae is supposed to preserve nomological necessity.) That this condition is stronger than the ontological requirement that whatever falls under the generalizations of T_1 should also fall under those of T_0 can be seen from the consideration that the latter could be satisfied even if bridge formulae expressed (not laws but) mere true empirical generalizations; but the requirement that nomological necessity be preserved under the substitutions that bridge formulae license would presumably *not* be satisfied in this case. (For further discussion, see Fodor, 1975.)

Second, if the present account is correct, there is an important sense in which *syntax* is preserved under reduction; on this view, the pertinent difference between a reduced generalization of T_1 and its reducing counterpart in T_0 lies just in the (nonlogical) *vocabulary* of the two formulae. To put this point the other way around, if we look only at the *form* of the sentences that constitute T_0 and T_1, disregarding such expressions as belong to the nonlogical vocabularies of the two theories, then there will not, in general, be any way of telling the sentences of T_0 and T_1 apart. Fundamentally, standard reduction just consists in substituting expressions in the nonlogical vocabulary of T_0 for expressions in the nonlogical vocabulary of T_1 under the equivalences which the bridge formulae specify.

It connects with the latter observation that the *point* of standard reduction (insofar as the point is not merely ontological) is primarily to exhibit the generalizations of T_1 as special cases of the generalizations of T_0. The idea is roughly this: events fall under the generalizations of T_1 by virtue of satisfying descriptions in the vocabulary of T_1. Reduction permits us to *re*describe these events in the vocabulary of T_0, hence to express their conformity to the generalizations of T_0. Since it is assumed that the generalizations of T_0 will normally hold in a domain that properly includes the domain of T_1 (e.g., physics is true of everything that psychology is true of, and physics is true of other things as well), progress in reduction should permit us to subsume phenomena under laws of increasing generality. But since it is said that the generality of laws is an index of the explanatory power of the theories that express them, progress in reduction turns out to be progress toward theories of increasing ex-

planatory power. The unity of science was, perhaps, initially construed as just a way of expressing the ontological claim that everything is the kind of thing that physics is about. But we can now see that the unity of science expressed an epistemological claim as well: the claim that physical explanation subsumes explanation in the special sciences.[4]

I want to consider the application of this picture of reduction to certain kinds of theories typical of current work in cognitive psychology. In order to do so, however, I shall first have to say something about the structure of such theories.

Computation

I am interested, for present purposes, in psychological theories that propose "computational" or "information flow" accounts of mental processes. Not all psychological theories do propose such accounts; indeed, not all mental processes provide appropriate domains for theories of this kind. Roughly, and to put the cart before the horse, mental states are computationally related only when they are related in content. Psychological theories of information flow model such relations of content by (a) providing a descriptive vocabulary in which the content of a mental state can be perspicuously represented, and (b) specifying transformations over formulae in that vocabulary that predict mental states and processes of the organism; in particular, its propositional attitudes. I have elsewhere discussed such theories at considerable length (Fodor, 1975), so in this paper I shall work largely from examples.

Consider learning. I suppose that cases of learning (or, in any event, cases of learning that . . ., where what fills the blank is approximately a sentence) are typically cases of environmentally determined alterations of epistemic state. In particular, what happens when someone learns that so and so is typically that (a) what he knows or believes changes, and (b) the change is a causal consequence (inter alia) of his transactions with his environment. So, for example, you can learn that Minneapolis is in Minnesota by looking at a map of Minnesota and noticing that Minneapolis is on it, or by hearing someone say (in a language one understands) that Minneapolis is in Minnesota, or by drawing the pertinent inference from the observation that Minneapolis is in the same state as St. Paul and

that St. Paul is in Minnesota, etc. In general, in such cases, certain things happen to one and, as a more or less direct consequence, what one knows or believes is altered in certain ways. I assume that "consequence" is to be construed causally here since, as far as I can see, no other way of construing it will suit the case.

But not every case of environmentally determined alteration in knowledge or belief is a case of learning. Suppose, for example, that someone were to invent a pill which, when swallowed, induces a mastery of Latin. One takes the pill and eo ipso acquires the relevant beliefs about what "eo ipso" means, how "cogitare" is conjugated, and so on. Moreover, the acquisition of these beliefs is, let us suppose, a causal consequence of taking the pill: the events of taking the pill and acquiring the beliefs fall, respectively, under the antecedent and consequent of a causal law, etc. Still, acquiring Latin by taking the pill is not *learning* Latin, any more than coming to speak the way aphasics do as a consequence of traumatic insult to Broca's area counts as learning aphasic. What is missing?

Intuitively, what is missing is this: the relation between what is acquired when one acquires Latin this way and the experiences that causally occasion the acquisition is, though nomological by hypothesis, notably arbitrary. (This contrasts with the case in which, e.g., one learns what "eo ipso" means by being told what "eo ipso" means, or by inducing its meaning from observations of occasions on which people say "eo ipso," etc.) A way to exhibit the arbitrariness is this: but for the hypothetical causal laws involved, one could imagine the situation to be reversed, so that it is insult to Broca's area that occasions the acquisition of Latin and swallowing the pill that induces aphasia. This situation seems no less gratuitous than the one we imagined initially. It is, in this sense, just an accident that the pills are connected with Latin rather than with aphasia; there is, as it were, nothing in a perspicuous description of what one learns when one learns Latin that connects it with what happens when one swallows a pill. But it is surely *not* just an accident that being told what "eo ipso" means is connected with learning what "eo ipso" means, or, for that matter, that it is English (and not Latin, Urdu, or aphasic) that children reared in English-speaking environments eventually learn to speak.[5]

Take another case. A man sees many gray elephants and, as a consequence of what he sees, comes to believe that elephants are gray. One wants to say there is a difference between this situation and one in which a man sees many gray elephants and, as a consequence of what he sees, comes to believe, say, that two is a prime number. What *kind* of difference? Well, not that the relation between seeing what he saw and coming to believe what he came to believe is causal, for we can imagine that to be true in *both* cases. Still, one wants to say, the first man learned (from his experiences) that elephants are gray, whereas the second many simply had certain experiences and came to believe that two is a prime number as a result of having had them. The relation between the second man's beliefs and his experiences is, in some important sense, arbitrary, whereas the relation between the first man's beliefs and his experiences, in the same important sense, is not. (It is, of course, connected with this that the experiences from which one can learn that so and so are often the experiences one can appeal to in justifying the belief that so and so.)

One more example, and then I shall try to say something about what the examples are examples of. A man sees many gray elephants and, as a causal consequence, comes to believe that elephants are gray. But, although each of the things he saw (the seeing of which contributed causally to the fixation of his beliefs that elephants are gray) was, in fact, a gray elephant, still what he *took* each of these things to be was, say, a very small, brown camel. Such a case is, of course, doubly grotesque; one wants to ask why a man should take elephants to be camels and why, having done so, he should come to believe that elephants are gray as a consequence of the putative camel-sightings. My point is that there need be no answer to these questions beyond adverting to the facts about the man's physical constitution and the way the world happens to impinge upon him. One can, in short, imagine a man so constructed and so situated that his experiences come to fix the right belief about the color of elephants by, as it were, the wrong route. But I think we should want to add that, prima facie, that sort of fixation of belief would not be learning.[6] For learning one needs a nonarbitrary relation (not just between the facts about the experiences and the content of the

beliefs they determine, but also) between the content of the beliefs and what the man *takes* the facts about the experiences to be.

To a first approximation, then: in paradigmatic cases of learning there is a relation of content between the belief that is acquired and the events that causally determine its acquisition. But this is a poor first approximation, because events do not, in general, have contents, although beliefs, in general, do. A better procedure is to relativize to descriptions and say that, in paradigmatic cases of learning, there is a relation of content between the belief acquired, under its theoretically pertinent description, and the events that causally determine the learning, under *their* theoretically pertinent descriptions. That is: one imagines an account of fixation of beliefs at large (hence of learning in particular) such that descriptions in some canonical language are assigned to the beliefs and to such organism/environment interactions as causally occasion the having of them. One further assumes that, under this assignment, it will sometimes turn out that there are relations of content between the former descriptions and the latter. Presumably all cases of learning will be cases of this kind. Indeed, one might take it to be a condition upon the adequacy of canonical psychological descriptions that this should, in general, be true.

But this is not good enough either. For, as we have seen, there could be cases in which experiences that are correctly described as experiences of gray elephants fix beliefs that are correctly described as beliefs that elephants are gray, yet the required relations of content do not hold between the experiences and the beliefs they fix. What is "transparently" an experience of gray elephants may be "opaquely" an experience of brown camels: if such an experience fixes a belief about elephant colors, the relations between the belief and the experience is, in the relevant sense, arbitrary (see note 6). This is a way of saying what psychologists have in mind when they emphasize the theoretical centrality of the *proximal representation* of the stimulus (as opposed to the distal stimulus per se) in any but the most superficial accounts of learning. (See the discussion in Dennett unpublished.)

We can fix this up as follows. We continue to reconstruct the notion of learning (as distinct from undifferentiated causal fixation of belief) in terms of content relations between experiences and

beliefs, both taken under their theoretically pertinent descriptions. But we construe theoretical pertinency as requiring an appropriate correspondence between the description the *psychological theory* assigns to the experience and the description the *subject* assigns to it. In effect, we construe theoretical pertinency of a description as requiring its psychological reality. If the subject internally represents what are in fact experiences of gray elephants as experiences of brown camels, then it is the latter description that enters into the psychological account of the relation between his experience and his beliefs. Descriptions of mental states are, in effect, read opaquely for the purposes of constructing such accounts.[7]

I have been considering the kinds of conceptual mechanisms a psychological theory will need if it is to preserve the distinction between learning that so and so and merely acquiring the belief that so and so. It appears, if the sketch I have given is even more or less correct, that at the heart of this distinction are certain constraints upon relations of content between beliefs and the experiences that fix them. It seems to follow that a psychology of learning will have to respect those constraints if it is to *be* a theory of learning; a fortiori, it will have to be able to represent the properties of mental states in virtue of which they satisfy such constraints—viz., the properties in virtue of which they have the content they do.

Now it may be thought that this sort of argument makes a great deal rest upon the preservation of a bit of ordinary language taxonomy; viz., on preserving the distinction between learning and mere causal fixation of belief. I want to emphasize, however, that that is not even slightly the sort of point I have in mind. I assume, rather, that the linguistic distinction probably corresponds to a fact *in rerum natura*; roughly, to the fact that there are generalizations that hold for learning but not for arbitrary cases of fixation of belief. I assume, moreover, that to state these generalizations we shall need to advert to the content of what is learned and to the content of the experiences that causally occasion the learning. To put the same claim the other way around, I assume that if we taxonomize mental states by their contents we shall be able to state general truths about them that we shall not be able to state otherwise; such truths, as, for example, that general beliefs tend to be fixed by experiences of their instances.[8]

I think it is, to put it mildly, very plausible that there *are* generalizations about mental states that hold in virtue of their contents, but I am not going to argue that claim here. Suffice it to emphasize its centrality not only in current approaches to the psychology of learning, but also in such adjacent fields as the psychology of perception, problem-solving, action, etc. In each case, theory construction proceeds by assigning canonical descriptions to mental states and by specifying functions from one such state to another. In each case, whether the theory represents a given mental state as falling under such a function depends on the canonical description the theory assigns to the state; the adequacy of the canonical description depends, in turn, on the accuracy with which it specifies the content of the mental state it applies to. A causal chain of mental states (e.g., the chains that run from experience to beliefs) thus gets a special sort of representation in this kind of theory: viz., a representation as a sequence of formulae related by content.

The explanatory power of such a treatment lies in its ability to predict the content of some mental states, given knowledge of the content of other, causally connected mental states. So, given a canonical representation of sensory contents, we should be able to predict the content of the percepts they give rise to. Given a canonical representation of a percept, we should be able to predict the memories it engenders; and so on, mutatis mutandis, wherever causally related mental states *are* related by content, viz., nonarbitrarily related. If, rather tendentiously, we take "coherent" to be the contradictory of "arbitrary," then the interest of computational psychological theories lies in their ability to explicate the principles according to which causally related mental states are also coherently related.

We have thus far been developing a picture of computational psychological theories as, in effect, treating causal relations among mental states as though they were derivational relations among formulae. It is, however, of prime importance to insist upon a point we encountered above: the interformulaic relations that such a theory articulates typically hold only insofar as the canonical descriptions of mental states are, as it were, construed opaquely.[9] So, for example, suppose it is true that general beliefs tend to be fixed by experiences of their instances. Then a theory of learning might tell us how

John's belief that elephants are gray is fixed by (what John takes to be) his experiences of gray elephants; e.g., given, as datum, that John took *n* of his experiences to be gray-elephant experiences, the theory might predict that John's belief that elephants are gray is fixed to degree *m*. But now 'elephant' is nonreferential in 'John believes elephants are gray' and in 'John took *e* to be a gray-elephant experience,' and 'gray' is nonextensional in those contexts. This is patently essential if the theory of learning is to be remotely plausible, since it seems clear that the very same experiences that fix the belief that elelphants are gray may be neutral to the belief that pachyderms reflect light of such and such a wavelength, and this may be true even though *for* elephants to be gray just *is* for pachyderms to reflect light of that wavelength. Whether a given belief is fixed by a given experience notoriously depends on how the belief and the experience are represented.

I have been arguing for the following contentions: on the one hand, information flow theories reconstruct content relations among mental states as computational relations among canonical descriptions; and on the other, because canonical descriptions specify the contents of mental states, they must be read opaquely. One way of putting the situation is this: if the general account of computational psychological theories I have sketched is right, then the possibility of constructing such theories depends on a certain approach to formulae embedded to verbs of propositional attitude in canonical psychological descriptions. Such formulae must be viewed as nonextensional but not as "fused."[10] To make this clear, I shall have to say a little about what fusion is supposed to be.

'Dog' is nonreferential in 'dogmatic.' But that is a bad way of putting it since "dog' (I mean the *word* 'dog' as opposed to the sequence of letters '$\widehat{\text{d' 'o' 'g}}$') does not so much as *occur* in 'dogmatic.' Similarly, according to the fusion story, 'elephant' is nonreferential in 'John believes elephants are gray,' because the word 'elephant' does not so much as occur in 'John believes elephants are gray.' Rather, 'believes-elephants-are-gray' is a fused expression, analogous to a one-word predicate or an idiom, so that the logical form of 'John believes elephants are gray' is simple F_{John}, indistinguishable from the logical form of, say, 'John is purple.' It is worth remarking, for later reference, that this is a two-step story. The nonreferentiality

of 'elephant' is explained by the assumption that it is not a term in 'believes elephants are gray,' and the denial of termhood is then rationalized by appeal to the notion of fusion. One can imagine alternative accounts on which 'elephant' is not construed as a term in 'believes elephants are gray' but on which verbs of propositional attitude are nevertheless *not* construed as fused with their objects. We shall return to this further on.

The present point, in any event, is this: fustion will certainly account for failures to refer; in something like the way being dead accounts for failures to be loquacious. But it is a kind of account that is not compatible with the development of psychological theories of the kind I have been describing. If "believes elephants are gray" is a fused expression, then a fortiori the canonical representation of John's mental state when he believes that elephants are gray bears no more intimate relation to the canonical representation of his mental state when he takes himself to sight a gray elephant than it does to the canonical representation of his mental state when, say, he takes himself to have sighted a brown camel.[11] Fusion is precisely a way of reading propositional attitudes as *not* exhibiting content, a fortiori as not exhibiting the relations of content.

Whereas, of course, the whole point of appealing to a notion of canonical psychological representation in the first place was to permit the development of, e.g., principles of fixation of belief that *are* sensitive to the way that mental states are related in virtue of content. So, in particular, the theory was to reconstruct the intuitive notion that there is a relation of content between experiences of elephants and beliefs about the color of elephants, and that the experiences tend to fix the beliefs in virtue of this relation. But if this whole strategy is to succeed, then it had better be that in canonical descriptions like 'believes elephants are gray' the object of 'believes' is somehow connected with the generalization *elephants are gray*, and in canonical descriptions like 'takes himself to see a gray elephant' the object of 'takes himself to see' is somehow connected with a singular statement about an elephant. Unless this condition is satisfied, we shall not be able to represent John's belief that elephants are gray *as* a general belief: a fortiori, we shall not be able to represent the fixation of that belief as falling under the principle that general beliefs tend to be fixed by experiences of

their instances. Conversely, if this condition *is* satisfied, then, by that very fact, it follows that the canonical representations of situations in which *a* has such and such a propositional attitude cannot, in general, be of the form *Fa*. In short, *we can have fusion or we can have computation, but we cannot have both*.

I take it to be the moral that any operation on canonical descriptions that has the effect of fusing the expressions they deploy will thereby deprive us of the very formal mechanisms on which the (presumed) explanatory power of computational psychological theories rests. My strategy in the rest of this paper will assume this is true. I shall argue (a) that the conditions on standard reduction could be satisfied even if canonical neurological representations of mental states are fused, hence (b) that the satisfaction of the conditions upon standard reduction does *not* guarantee the subsumption of psychological explanation by neurological explanation. The form of argument is thus that fusion is a sufficient condition for loss of explanatory power and that standard reduction is compatible with fusion, hence the success of standard reduction would not, in and of itself, ensure that the kinds of explanations that computational psychological theories yield can be reconstructed in the vocabulary of the neurological theories that reduce them. One can look at such an argument as showing that there is something wrong with the standard notion of reduction (since standard reduction turns out to be compatible with loss of explanatory power). Alternatively, one can hold to the standard notion of reduction and abandon the claim that explanation in a reduced science is subsumed by explanation in the reducing science. My own inclination, for present purposes, is to take the former line and strengthen the constraints on the neurological reducers of psychological theories; I shall return to this in the last section.

Computation and Reduction

I remarked above that, in paradigm cases of classical reduction, mapping the sentences of T_1 onto sentences of T_0 is primarily a matter of replacing items in the vocabulary of the former with items in the vocabulary of the latter, such replacements being mediated by the lawful coextensions (or identities; the distinction is not germane to the present argument) that bridge laws express. But it should

now be clear that this will *not* be the case in the reduction of (computational) psychological theories to, say, neurology. For, as we have seen, computational psychological theories contain canonical descriptions that make serious use of a formulae in which sentences are embedded to opaque verbs ("serious" in the sense that the generalizations the theory articulates depends critically on the form and vocabulary of such embedded sentences); whereas at least on the usual assumptions about neurology (and, a fortiori, about physics) those sciences do not employ descriptions of that kind. So the reduction of psychology to neurology (unlike, say, the reduction of meteorology to mechanics) involves alteration of the syntax of the reduced formulae, and it is easy to see from examples that the effect of such alteration will typically be the fusion of expressions that specify the objects of propositional attitudes.

Consider the reduction to neurology of a psychological theory containing the formula 'John believes elephants are gray.' Given the usual assumptions, there will be a sentence of neurological theory (say 'John is *N*') such that 'John believes elephants are gray iff John is *N*' is nomologically necessary. So let us suppose that 'John believes elephants are gray' reduces to 'John is *N*,' and similarly, mutatis mutandis, for 'John takes *e* to be a gray elephant experience,' which comes out under reduction to be, let us say, 'John is *M*.' Given this much, the classical constraints upon the reduction of psychology to neurology are satisfied insofar as they apply to these two sentences. And if, as we may suppose, 'John's being M brings about John's being N' instantiates a causal law, then we have a causal explanation, in the vocabulary of neurology, of the contingency of John's belief about elephant colors upon John's experiences of colored elephants. So far, so good.

Except, of course, that fusion has already occurred (taking, as the criterion of fusion, the failure of canonical—now neurological—descriptions to specify the content of the mental states they apply to). To see this, imagine that, by some or other causal quirk, not only (putative) experiences of gray elephants, but also some experiences that do not have contents (like swallowings of pills) or some experiences that have the "wrong" contents (like putative sightings of camels) happen to be causally sufficient for fixing beliefs about the color of elephants. Then, presumably, there will be

an expression *E* that is (a) in the language of neurology, (b) such that 'John is *E*' is true if John swallows the pill (takes himself to sight the camel), and (c) such that 'John's being *E* brings about John's being *N*' is *also* an instance of a causal law. That is, once we go over to neurological descriptions, there need be nothing to choose between the way the theory represents the case in which John's coming to believe elephants are gray is consequent upon his sighting gray elephants and the case in which John's coming to believe elephants are gray is consequent upon his swallowing blue pills. Looked at formally, this is due to the fact that reduction permits the fusion of 'believes elephants are gray' (and/or 'takes himself to sight a gray elephant'), where the mechanism that accomplishes fusion is the substitution of some (possibly elementary) neurological expression for a psychological expression in which a verb of propositional attitude has scope over a formula that specifies the content of a propositional attitude. Looked at substantively, what has been lost is a representation of the relation of content between beliefs about elephants and elephant-experiences. If, then, there are generalizations that hold of mental states in virtue of the content relations between them (if, for example, there are generalizations that relate the content of beliefs to the content of the experiences that fix them), then the conditions on reduction may be satisfied even though such generalizations (statable by assumption in the psychological vocabulary) are not statable in the vocabulary of the reducing science. In short, insofar as there is any explanatory power to be gained by resort to a computational psychology, reduction is in danger of losing it for us.

I had better, at this point, make as clear as I can what I am *not* claiming. To begin with, I do not deny that there could be a truth of neurological theory that applies to exactly the cases in which say, a general belief is fixed by its instances. On the contrary, if the truths of psychology are to follow from the truths of neurology plus bridge laws, there had *better* be a neurological state necessary and sufficient for having any given belief, and a neurological state necessary and sufficient for having any given belief-fixing experience; and the neurological theory had better say (or, anyhow, entail) that states of the latter kind are causally sufficient for bringing about states of the former kind. The difficulty is, however, that since the contents

of the beliefs and experiences presumably will not be specified by their *neurological* descriptions, it is only when we are given their *psychological* descriptions that we will be able to predict the contents of the beliefs *from* the contents of the experiences. I am saying, in effect, that beliefs and experiences reduce to neurological entities, but the contents of beliefs and experiences—the things that our beliefs and experiences relate us to—do not reduce to anything; psychological representations of content simply fuse under neurological description of mental states. So, to put it rather misleadingly, although neurology can, in principle, say anything that needs to be said about the contingency of beliefs upon experiences, it has no mechanisms whatever for talking about the contingency of the contents of beliefs upon the contents of experiences. Yet there are, it appears, such contingencies, and there are interesting things to be said about them.

It might nevertheless be held with some justification that this line of argument is unfair, if not to the spirit of standard reductionsim, then at least to the letter. For, on the reductionist view, T_0 will not reduce T_1 unless all the consequences of T_1 are consequences, not of T_0 alone, but of T_0 together with the bridge laws. Now, if there are bridge laws of the form: (x) (x has a general belief [of the appropriate content] iff N_x), and (y) (y has a belief-fixing experience [of the appropriate content] iff M_y), and if 'M brings about N' expresses a law of neurology, then neurology together with the bridge laws *does* entail whatever psychology entails about the fixation of beliefs by experiences.

Still, the present case is quite unlike what the classical reduction paradigm envisions. True, in the standard examples, we need not only T_0 but also the bridge laws to recover the entailments of T_1. But that is only for the relatively uninteresting reason that the bridge laws provide access to the (nonlogical) *vocabulary* of T_1, which is, by assumption, not included in the vocabulary of T_0. (Chemistry can entail "H_2O is wet," but only chemistry plus the bridge laws can entail "Water is wet.") Whereas the curious thing about the psychology/neurology case is that here the bridge laws provide access not just to the vocabulary of the reduced science but also to an *explanatory construct*—content—for which the reducing science offers no counterpart. Specifically, we shall need the bridge laws to

unpack the fused objects of verbs of propositional attitude if, as I have argued, fusion deprives us of the appropriate theoretical mechanisms for specifying the domains in which generalizations about cognitive processes hold.

In short, we are back where we started. I have argued, *not* that the classical constraints upon reduction cannot be met in the psychology/neurology case, but rather that they fail to provide sufficient conditions for the subsumption of psychological explanations by neurological explanations. And examination of that case has shown precisely that it is possible for a pair of theories to meet the classical constraints (T_0 plus the bridge laws entails whatever T_1 does) even though intuitively (and by the fusion test) the explanations of T_1 are not subsumed by T_0. Requiring that T_0 together with the bridge laws yield the entailments of T_1 does not ensure that the explanations available in T_1 have counterparts in T_0, Q. E.D.

I have been issuing caveats. Here is another: the present argument is not that reduction *must* lose the advantages that psychological models gain; only that it *can* do so compatibly with the satisfaction of such conditions, ontological and methodological, as standard views of reduction impose. This is due to the fact that nothing in those conditions prohibits fusion as the consequence of reduction. On the contrary, in the absence of further constraints upon reduction, fusion would be its *natural* consequence, as can be seen from the following.

Reduction required, in effect, that for every psychological state of John's there exist a coextensive (or token-identical) neurological state of John's, and that every psychological sentence that attributes the former state to him should be replaced by a neurological sentence that attributes the latter state to him. But consider again the sentence, 'John believes elephants are gray.' The shortest stretch of that sentence that can be construed as expressing a state (property, etc.) of *anything* is surely 'believes elephants are gray,' since, in particular, in this sentence the occurrence of elephants is nonreferential and 'are gray' does not express a property of elephants. In short, if we are to substitute neurological-state expressions for psychological-state expressions, the natural choice is to make the substitution in the frame: ⌜John . . .⌝, i.e., to substitute simultaneously for the verb of propositional attitude and its object. And,

since the classical construals of reduction do not do anything like requiring that the content of propositional attitudes can be specified by neurological representations of mental states, the consequence of substitution in this frame is likely to be, precisely, fusion.

I have a strong suspicion that this chapter would do well to stop here. For I suspect the moral just drawn is essentially the right one: reduction will probably require fusion, and fusion will entail the loss of the explanatory power that computational psychological theories are constructed to obtain. If this is true, it suggests that we will have to be very much more pluralistic about scientific explanation than classical views of the unity of science supposed. In particular, nothing in the discussion has jeopardized the ontological claim that mental states are neurological states; on the contrary, the whole argument can be run on the standard assumptions of the mind/body identity theory. But what turns out not to be true is that explanation in a reduced science is invariably subsumed by explanation in its reducer. Rather, we shall have to say something like this: Mental states have canonical psychological descriptions in virtue of which they fall under the generalizations expressed by computational principles, and they have canonical neurological descriptions in virtue of which they fall under the generalizations expressed by causal laws. Quite possibly there never will be a state of science which we can, as it were, do neurology *instead of* psychology because, quite possibly, it will never be possible to express in the vocabulary of neurology those generalizations about relations of content that computational psychological theories articulate. Pyschologists have lots of things to worry about, but technological unemployment is not likely to be one of them.

It may, however, be worth forging on. I want to sketch, very rapidly and incompletely, a way that psychology and neurology might turn out so as to make possible reduction without fusion. I am not going to defend the claim that either psychology or neurology *will* turn out that way. My primary interest is still just to make clear how much more than correlation (or contingent identity) of psychological and neurological states the substantive reduction of psychology to neurology would require.

Reduction without Fusion

What we have said so far amounts to this: we want a psychological theory that at least provides canonical descriptions of mental states, and we want canonical descriptions to reconstruct the contents of the mental states they apply to. Insofar as such a theory is formalized, its generalizations will apply to mental states in virtue of features of their canonical representations. Such a theory should therefore suffice to represent the causal sequences that constitute the mental life of an organism by sequences of transformations of canonical representations. To contemplate the substantive reduction of computational psychology is, in effect, to suppose that such theories can operate solely with neurological constructs. To put this last point slightly differently, it is to suppose that the descriptions in virtue of whose satisfaction psychological states fall under principles of computation are descriptions in the same vocabulary as those in virtue of whose satisfaction psychological states fall under neurological laws. The question is whether we can imagine a reduction of psychology to neurology that makes this true.

We have seen that the basic methodological problem is to find a way of representing the contents of mental states that avoids recourse to fusion while doing justice to the nonreferentiality of terms occurring in typical psychological contexts. There is a classical proposal here that, as far as I can see, may well point in the direction in which we ought to look: take verbs of propositional attitude to express relations between organisms and *formulae*. In particular, on this view, to believe that elephants are gray is to be related, in a certain way, to some such formula as 'elephants are gray'; to take oneself to see a gray elephant is to be related in a certain (different) way to some such formula as 'I see a gray elephant,' etc.[12]

There is a well-known difficulty with this suggestion, but I think it has been overplayed: viz., believing that elephants are gray *cannot* be being related (in whatever way) to the formula 'elephants are gray,' since, if it were, it would presumably follow that monolingual English speakers cannot have the same beliefs as, say, monolingual French speakers. And it would also presumably follow that infraverbal organisms (cats, dogs, and human infants, inter alia) can have no beliefs at all.

The most that this objection shows, however, is not that believing cannot be being related to a formula, but only that, if it is, then all organisms that can have shared beliefs must have some shared language.[13] I am convinced, for reasons I have elaborated elsewhere (Fodor, 1975), that we would be well advised to take that suggestion seriously; in fact, that it is quite impossible to make sense of the notion of a computational psychology unless some such suggestion is endorsed. The idea is, roughly, that all organisms that have a mental life at all have access to some system of internal representations; that insofar as the mental life of organisms is homogeneous (e.g., insofar as people and animals, or, for that matter, people and machines, instantiate the same psychology) there must be corresponding homogenieties between their internal representational systems; and that a major goal of information flow theories must be to characterize this system of representations and provide necessary and sufficient conditions for the having of propositional attitudes by reference to relations between organisms and the formulae of the system. On this view, for example, to believe that elephants are gray is to be in a certain relation to whatever internal formula translates the English sentence, 'Elephants are gray'; if there is a content relation between that belief and certain of the experiences that are causally responsible for fixing it, then that relation is expressed by generalizations defined over whichever internal representations are implicated in the having of the belief and experiences. In effect, there is a language of thought, and content relations among propositional attitudes are to be explicated as relations among formulae of that language.[14]

Correspondingly, the canonical representations deployed by a computational psychology are assumed to contain structural descriptions of internal formulae. Mental states fall under the generalizations articulated by psychological theories because they satisfy their canonical representations, so John's believing that elephants are gray makes true a certain psychological sentence; viz., a sentence of the form $R_{\text{John}}\,SD$. In that sentence, SD is the structural description of an internal representation (in particular of the internal representation which translates "elephants are gray") and R is a relation between John and that internal representation (in particular,

whichever relation to an internal representation is nomologically necessary and sufficient for believing what it expresses).[15]

We are so far from having a developed cognitive psychology that it is hard to give untendentious examples. But consider the propositional attitude *remembering*, and suppose (for once *not* contrary to fact) that psychology acknowledges a relation of *storing* that holds between organisms and internal representations. Then the following might be among the sentences that psychology entails: ⌜John remembers (the fact) that elephants are gray iff John stores (the formula SD⌝, where what substitutes for 'SD' is the structural description of the internal translation of 'elephants are gray.' Note that the biconditional is extensional for the object of 'stores.' That is, it remains true whatever name of the internal formula one substitutes for 'SD.'

However, structural descriptions (unlike other kinds for names) play a special role in this sort of theory, and this connects with the fact that, strictly speaking, structural descriptions are not *names* at all. What they are, of course, is descriptions. So, suppose that 'Alfred' is a name of the internal formula SD. Then, although we preserve *truth* if we substitute 'Alfred' for a structural description of SD in psychological sentences containing the canonical representation of SD, we *do not*, in general, preserve canonicalness. A canonical representation of a mental state must specify its content. We get such a specification (ceteris paribus) insofar as the canonical representation of a mental state contains the structural description of an internal formula, but we do not get it (ceteris paribus) when it contains a noncanonical name of that formula like 'Alfred.' The general idea is that internal representations (like, for that matter, English sentences) express the content they do because they satisfy the structural descriptions they do. Structural descriptions specify those properties of a formula that determine its syntactic and semantic behavior—those properties by virtue of which a formula constitutes an expression in a language.

The assumption that canonical psychological representations typically contain structural descriptions of internal formulae allows us some of the advantages of fusion theories without their most obvious vices. In particular, if the canonical counterpart of 'John

believes elephants are gray' is of the form $R_{\text{John}, SD}$, it is not surprising that the canonical counterpart of 'elephant' fails to refer to elephants when it occurs within the scope of the canonical counterpart of 'believes.' Roughly, the present account agrees with the fusion story in holding that 'elephant' is not a term in 'John believes elephants are gray.' But it provides a different rationale for the denial of termhood. Since the immediate epistemic objects of propositional attitudes are taken to be formulae, the syntactic objects of verbs of propositional attitude are taken to be structural descriptions of formulae; the word 'elephant' is not a term in 'believes elephants are gray,' but the name of that word is.

The difference between this view and the fusion story should not be minimized. Structural descriptions are unfused expressions: qua names, they purport to refer, and qua descriptions, they purport to determine their referents in virtue of the properties of their referents. Correspondingly, verbs of propositional attitude are construed relationally on the present account; in particular, they express relations between organisms and the referents of structural descriptions; i.e., relations between organisms and formulae of the internal representational system. Such relations are ontologically kosher. No fusion theory can make that statement.

Suppose, then, that canonical psychological representations turn out to contain structural descriptions of internal formulae. Is there any way of reducing this sort of psychology to neurology without committing fusion at the point of reduction? If the argument we have been pursuing is correct, that is what the issue about the possibility of substantive reduction—reduction without loss of explanatory power—boils down to in the case of cognitive psychology.

I suppose the answer goes like this: substantive reduction would at least require (1) that token computational processes turn out to be token neurological processes (storing a formula turns out to be a neurological process, etc.); (2) token internal representations turn out to be token neurological states (a token internal representation that translates 'elephants are gray' turns out to be some neurological configuration in, roughly, the way the above sentence token is a configuration of ink marks on this page); and (3) canonical names of internal formulae (viz., their structural descriptions) are specifiable in the vocabulary of neurology.

I take it that (1) and (2) are just consequences of applying the usual ontological conditions upon reduction to the special case of psychological theories that acknowledge internal representations. They do not, that is, distinguish substantive reduction from standard reduction. It is (3) that does the work. In effect, (3) requires that the canonical *neurological* description of a mental state (of *a*'s) be of the form R_a, *SD*, (and not, for example, of the form R_a, *Alfred*). So the question that has to be faced is: what would have to be the case in order for (3) to be satisfied? Heaven knows, I am unclear about how that question should be answered, but what I *think* it comes to is this: for psychology to be substantively reducible to neurology, it must turn out that neurological entities constitute a code, and that the canonical neurological representation of such entities specifies the properties in virtue of which they constitute formulae in that code. Since the properties in virtue of which a formula belongs to a code are the ones in virtue of which it satisfies its structural description, and since the properties in virtue of which a formula satisfies its structural description are the ones in virtue of which it has the content it has, we can summarize the whole business by saying that neurology will not reduce psychology unless neurological descriptions specify the content of internal formulae. (Compare the standard view, in which what specifies the content of a mental state is its canonical neurological representation *together with the relevant bridge laws*, and in which the specification is couched in the vocabulary of the reduced rather than the reducing science.)

I have gone about as far as I can, but it is worth remarking that the notion that some neurological states do constitute a code is not exactly foreign to the speculative literature on brain and behavior. So, one might suppose, neurons are relays and canonical names of neurological states are specifications of levels of neural excitation. For this to be true, it would have to turn out that to specify the state of excitation of a set of neurons is to fix the content of a token formula, just as specifying the structural description of an English sentence fixes the content of its tokens. I do not have the foggiest idea whether anything like that *is* true, but the following, at least, is clear: if neurological representations specify those properties of states of the central nervous system in virtue of which they

constitute formulae belonging to a code, then the descriptions that such states receive in sciences still more basic than neurology almost certainly do not. (Think what a particle description of a token neurological state—or, for that matter, a token English sentence—would actually look like; then try to imagine specifying, in that vocabulary, such properties as those in virtue of which a sentence token like 'elephants are gray' is content-related to a sentence token like 'there's a gray elephant.' To specify such relations, we need, e.g., notions like 'quantifier' and 'general term.' Is it plausible that such notions should be expressible in the vocabulary of particle physics? The more reason we have for thinking that neurology might substantively reduce psychology, the less reason we have for thinking that physics might substantively reduce neurology.[16]

We have come quite a long way from the suggestion that what we need to reduce psychology to neurology is just correlation (or token identity) of psychological and neurological states. And, as we anticipated at the start, it is the intentionality of psychological predicates that primarily confounds that suggestion. On the one hand, terms in formluae embedded to psychological verbs are typically nonreferential, but, on the other, it is precisely such formulae that express the contents of mental states; and, in the theoretically interesting cases, mental states are related in virtue of their content. Insofar as reduction leads to fusion, it thereby results in the failure to represent the generalizations about mental life that structure such relations. But these generalizations are, as we remarked above, involved in the very rationality of mental life. (It is constitutive of the rationality of John's beliefs about the color of elephants that they are fixed by, e.g., his elephant-sightings and not, e.g., by swallowing pills, or sighting camels, or having his cortex surgically rewired.) Small wonder that antireductionists have often held that to replace psychological explanations by neurological explanations is to lose precisely what a theory of the mind ought to be about. Given the standard notion of reduction, this objection seems to me entirely pertinent.

There is, however, an undefended premise in this whole argument, and I had better say something about it before I stop. I have argued that neurological representations will, quite possibly, fail to provide the appropriate format for such generalizations as hold in virtue of

content-relations of mental states. But it might be replied that there are, in fact, no such generalizations; that the distinction between believing elephants are gray because of all those gray elephants and believing elephants are gray because of all those blue pills is not a distinction that a scientifically disciplined theory of mental states would recognize. Of course, we, pretheoretic as we are, like to draw such distinctions; and of course, insofar as they come to anything at all, there will be distinctions between causal mechanisms corresponding to (viz., coextensive with) what we take to be distinctions among relations of content. But a theory of content-relations per se would not be formulable in a first-class conceptual system. What counts, in such a system, is those descriptions under which events (including mental events) instantiate the laws of basic science. And nothing else counts.

Underlying this objection (if I understand it correctly) is the observation that no behavioral or neurological (or physical) description of an organism will uniquely determine an assignment of propositional attitudes to that organism. Any such assignment, however plausible in the light of the behavioral and neurological data, is to that extent an interpretation that we place upon the physical facts and hence not something to be mentioned in even the most exhaustive catalogue of the physical facts themselves.[17] (That assignments of propositional attitudes are interpretations of the physical facts would itself not be mentioned in such a catalogue; interpretation is not a physical category either.) But if there are not any facts about propositional attitudes, then a fortiori there is not the fact that internal formulae are nomologically implicated in the having of propositional attitudes; this is also true however plausible it turns out to be to treat neurological states as tokens in a code and however much such a treatment seems to rationalize the behavioral observations. We need internal formulae to account for propositional attitudes, and we need propositional attitudes if we are to represent such facts as that organisms act out of their beliefs and utilities. But if there are no such facts the whole pattern of explanation is otiose.

Whatever else there is to be said on this issue, however, it is essential to distinguish it from the question of the substantive reducibility of psychology. For the latter is, by hypothesis, an empirical matter, whereas the whole point about the underdetermination of mental

ascriptions by physics is that, if it is true at all, it is true *however* the physics and the psychology turn out. Suppose our evidence for treating a certain neurological state as a token of a certain linguistic type were as good as our evidence for treating 'it's raining' as a token of the English type *it's raining*. That would not advance the case one jot since, on the present view, the assignment of token inscriptions to English sentences is *also* just a gloss upon the physical facts. There must be indefinitely many ways of associating objects of the physical form "it's raining" with linguistic objects *salve* the totality of physically characterizable facts about the organisms which produce such tokens.

On the other hand, there is nothing in this line of argument that stops our evidence for the linguistic analysis of token neurological states from being as good as—indeed, of the same kind as—our evidence for the linguistic analysis of English inscriptions, and, skeptical worries to one side, it is hard to believe the latter evidence is other than pretty good. There is thus room for a program of empirical research this side of skepticism: show that *if* there is good reason for treating (some) inscriptions as linguistic tokens, then there is *equally* good reason for treating (some) neurological states as linguistic tokens. It would, in short, be enormously impressive to show that neurological objects satisfy relevant necessary conditions for interpretation as a code, even if it turns out that nothing could show *which* code they (or anything else) belong to.[18] Such a demonstration would be tantamount to the substantive reduction of computational psychological theories. My point throughout has been that nothing less will do.

Notes

1. When I speak of intentionality, I shall usually have two related facts in mind. First, that psychological states (including, specifically, propositional attitudes) are typically individuated by reference to their *content*; second, that expressions that occur in linguistic contexts subordinate to verbs of propositional attitude are typically nonreferential. It is notoriously hard to say how, precisely, the first of these facts is to be construed or what, precisely, the relation between the two facts is. Some of the discussion in this paper is tangent to those issues, but I shall dodge them whenever I can. I shall not, in particular, be attempting anything so ambitious as a general theory of intentionality.

2. Quine (1960) and Dennett (1971) are perhaps the best examples of influential physicalists to whom this charge does *not* apply.

3. The sketch of classical reductionism I have just given is very inadequate from a

number of points of view. Nothing in the following discussion will exploit its inadequacies, however, and it would take considerable space to do justice to the details of the proposal.

4. I think it was a pervasive and characteristic error in positivistic thinking to infer the unity of science from the unity of the subject matter of science; viz., the epistemological thesis from the ontological one. However that may be, it is easy to find passages in the positivist literature in which the former doctrine is espoused in no uncertain terms. Thus Hempel (1949, p. 382) wrote, "The division of science into different areas rests exclusively on differences in research procedures and direction of interest; *one must not regard it as a matter of principle. On the contrary, all the branches of science are in principle of one and the same nature; they are branches of the unitary science, physics.*" (Emphasis Hempel's.) I should add that Hempel has since disavowed many of the ideas in that paper, and I do not intend to suggest that the passage I have quoted is indicative of his present views.

5. The experiences that lead to the acquisition of the rules of English are, normally, observations of utterances that formally instantiate the rules of English; there is, in that sense, a connection between the content of what is learned when one learns English and the content of the experiences that occasion the learning, and it would be a grievous error for a theory of learning to miss that connection. The equal and opposite error would be to try to parlay such connections of content into conceptual necessities, as, I think, some "ordinary language" philosophers have been inclined to do. It is not logically necessary in any useful sense of "logic" that hearing English is normally causally sufficient for speaking English; surely there are possible worlds in which it is normally causally sufficient for speaking Urdu.

6. The "prima facie" is important. For, of course, one can imagine a case in which someone knows something from which it follows (deductively or plausibly) that if small camels are brown then elephants are gray. In this case, though not only in this case, the fixation of the belief that elephants are gray by putative experiences of brown camels need not be arbitrary and might (at least to that extent) count as learning. A serious attempt to distinguish between learning and mere causal fixation of belief would, in short, need to work with a far deeper notion of "nonarbitrariness" than the examples so far might suggest. I am not, however, trying to draw such a distinction here; only to give a rough indication of the direction in which it lies. (The present case is, by the way, just a "Gettier example" transferred from "knows" to "learns", see Gettier, 1963.)

7. I do not suppose we can generally identify the internal representation that the subject assigns the stimulus with the representation he would (or could) supply if asked. Nor do I suppose this point needs, by now, to be argued.

8. I have used this example throughout as a paradigm of a generalization about relations between mental states that appears, prima facie, to be statable only by reference to their contents. I like this example because it is so pedestrian; it is hard to see how any psychology of learning could fail to have some such principle among its tenets. But I do not want to suggest that such examples are hard to come by. On the contrary, they are the cognitive psychologist's stock in trade. The contingencies that cognitive psychologists try to articulate are precisely those in which the contents of mental states are dependent or independent variables, or both.

It is worth emphasizing that the present account takes a view of the domain of cognitive psychological explanation different from the one that dominates the philosophical literature. It has become a sort of dogma that explanations that appeal to the contents of mental states can function only when the states are "rationally" related. The extreme form of this doctrine is that such explanations have literal application only to an ideally rational entity. (For versions of this story, see Dennet [1971], Quine [1960], Davidson [1970].)

Now it presumably *is* true that rationally related mental states are so related in virtue of their content. But there are plenty of cases of plausible psychological generalizations that hold for mental states that are content-related but *not* rationally related, not even on the most inflationary construal of "rational." Association by similarity (if there is such a process) is a simple and venerable example. Or consider the well-supported psycholinguistic generalization that sentences of negative import are harder to understand and to remember than corresponding affirmatives. Such a generalization would seem, on the face of it, to apply to propositional attitudes in virtue of features of their content. But there is no obvious sense in which the mental processes it envisages are usefully sigmatized as rational. On the contrary, an ideally rational creature would presumably never forget, or misunderstand, anything at all. (For further discussion, see Fodor, forthcoming.)

9. This is to put it very roughly. We shall see what it comes to in more detail as we go along.

10. The notion that the (syntactic) objects of verbs of propositional attitudes are (semantically) fused expressions (or, what comes to pretty much the same thing, that verbs of propositional attitude must be read "non-relationally") is one that a number of philosophers have flirted with, either in the context of discussions of Leibniz's Law problems about the mind/body identity theory or in the (intimately related) context of worries about the ontological commitments of psychological ascriptions. It is not clear to me that anybody actually holds the fusion theory of propositional attitudes, but for discussions in which that option is contemplated, see Quine (1960), Nagel (1965), and Dennett (1971). The term "fusion" is borrowed from Dennett.

11. For that matter, the canonical counterpart of 'John takes himself to see a gray elephant' will have no relation (other than the reference to John) to the canonical counterpart of 'John takes himself to see an elephant,' even though (one might have thought) taking oneself to see a gray elephant *is* taking oneself to see an elephant (inter alia). If, in short, the theory wants to represent these states as connected, it will have to do so by specific stipulation; e.g., by taking some such principle as '*x* takes-himself-to-see-a-gray-elephant → *x* takes-himself-to-see-an-elephant' as a nonlogical axiom. Fused representations are, to put it mildly, semantically imperspicuous.

12. By contrast, seeing a gray elephant is not being related to a formula, it is being related to an elephant. The present proposal concerns the construal of psychological verbs whose (syntactic) object is read opaquely. I have nothing at all to say about the notoriously difficult question of how to construe such verbs when their objects are read transparently.

13. In fact, it shows less, since it would do, for these purposes, if every organism had a language and the languages were intertranslatable insofar as the mental states of organisms overlap. It is no news, of course, that issues of translation and issues of the proper treatment of propositional attitudes tend to merge.

14. I discover, very belatedly, that an account in some respects quite like this one was once proposed by Sellars (1956). Sellars's work seems remarkably prescient in light of (what I take to be) the methodological presuppositions of contemporary cognitive psychology.

15. For the benefit of those keeping score, we note that the following pieces are now in play. There are (a) internal representations. These are formulae in an internal language, and it is assumed that they are the immediate objects of propositional attitudes. In particular, nomologically necessary and sufficient conditions for the having of a propositional attitude are to be formulated in terms of (presumably computational) relations that the organism bears to internal representations.

There are also (b) structural descriptions of internal representation. These are formu-

lae in the vocabulary of an (ideally completed) psychological theory. Structural descriptions are canonical names of internal representations (see text below). A propositional attitude has the content that it does because the internal representation that constitutes its immediate object satisfies the structural description that it does.

There are also (c) English sentences like 'Elephants are gray.' For heuristic purposes, we use such sentences to form definite (but noncanonical) descriptions of internal representations. We do this because we do not know what the structural descriptions of internal representations are like. That is: ideally completed psychological theories refer to internal representations via their structural descriptions. *We* refer to them as, e.g., the internal representation that translates 'elephants are gray.' We do so faut de mieux.

Finally, there are (d) structural descriptions of English sentences. Structural descriptions of English sentences specify the properties in virtue of which *they* have the content that they have. Roughly, structural descriptions are specifications of sentence types couched in an ambiguity-free notation. If what most psycholinguists now believe is true, the structural descriptions of English sentences must themselves function as internal representations. For, on current views, structural descriptions are normally (among) the ones that speakers intend their utterances to satisfy and that hearers recover in the course of construing the utterances of speakers. That is: in the theoretically interesting cases, the internal representation of an English sentence *is* its structural description.

16. I want to reemphasize that I am *not* denying that the (putative) neurological sentence tokens will satisfy *some physical descriptions or other*, just as the present sentence token satisfies some physical description or other. The question is whether their *physical* descriptions will turn out to be construable as *structural* descriptions which individuate the sentence types that the tokens belong to. (The corresponding question for natural language tokens is, approximately: does a formant analysis of an utterance represent its logical syntax. To which the answer is, of course, "resoundingly, no!")

17. Of course the merely notional status of propositional attitudes does not *follow* just from the observation that mental ascriptions are not entailed by physical ascriptions. what follows from that is only that behaviorism is false and physicalism is not better than contingently true. To get the result that propositional attitudes are fictions one needs to add some such premise as that their ascription would not be justified *unless* it followed from physics. I do not know how, precisely, such a premise would be formulated or how it could be defended.

It is worth mentioning, by the way, that the logical independence of mental and physical statements goes in both directions and has supported dubious arguments both ways. It used to be claimed that tables and chairs are notional on the grounds that physical object statements are not entailed by statements about percepts. Ho hum!

18. For example, internal representations must be at least as differentiated as the contents of propositional attitudes if they are to play the role that we have cast them for in individuating the contents of propositional attitudes. This is a strong condition; one that is not satisfied, e.g., by English orthographic sequences, since the latter do not constitute an ambiguity-free notation.

References

Davidson, D. Mental events. In Forster, L. & Swanson, J. (Eds.), *Experience and theory*. Amherst, Mass.: University of Massachusetts Press, 1970.

Dennett, D. Skinner skinned, a diagnosis of B. F. Skinner's Central Error. Tufts University. Unpublished.

Dennett, D. Intentional systems. *Journal of Philosophy*, 1971, 68, (4), 82-106.

Fodor, J. A. *The language of thought*. New York: Thomas Y. Crowell, 1975.

Fodor, J. A. Three cheers for propositional attitudes; a reply to Dennett's "Intentional Systems." Forthcoming.

Gettier, E. Is justified true belief knowledge? *Analysis*, 1963, 23, 121-123.

Hempel, C. G. Logical analysis of psychology. In Feigel, H. & Sellars, W. (Eds.), *Readings in philosophical analysis*. New York: Appleton-Century Crofts, 1949.

Nagel, T. Physicalism. *The Philosophical Review*, 1965, 74, 339-356.

Place, U. T. Is consciousness a brain process? *British Journal of Psychology*, 1956, 47, 44-50.

Quine, W. V. *Word and object*. Cambridge, Mass.: M.I.T. Press, 1960.

Sellars, W. Empiricism and the philosophy of mind. In Feigel, H. & Scriven, M. (Eds.), *Minnesota studies in the philosophy of science*, Vol. I. Minneapolis, Minnesota: University of Minnesota Press, 1956.

Smart, J. J. C. Sensations and brain processes. *Philosophical Review*, 1959, 68, 141-156.