WHAT PSYCHOLOGICAL STATES ARE NOT[1]

I

AS FAR as anyone knows, different organisms are often in psychological states of exactly the same type at one time or another, and a given organism is often in psychological states of exactly the same type at different times. Whenever either is the case, we shall say of the psychological states of the organism(s) in question that they are *type identical.*

One thing that currently fashionable theories in the philosophy of mind often try to do is characterize the conditions for type identity of psychological states. For example, some varieties of philosophical behaviorism claim that two organisms are in type-identical psychological states if and only if certain of their behaviors or behavioral dispositions are type identical. Analogously, some (though not all) varieties of physicalism claim that organisms are in type-identical psychological states if and only if certain of their physical states are type identical.[2]

In so far as they are construed as theories about the conditions for type identity of psychological states, it seems increasingly unlikely that either behaviorism or physicalism is true. Since

[1] A number of friends and colleagues have read earlier drafts. We are particularly indebted to Professors Richard Boyd, Donald Davidson, Michael Harnish, and Hilary Putnam for the care with which they read the paper and for suggestions that we found useful.

[2] If physicalism is the doctrine that psychological states are physical states, then we get two versions depending whether we take "states" to refer to types or to tokens. The latter construal yields a weaker theory assuming that a token of type x may be identical to a token of type y even though x and y are distinct types. On this assumption, type physicalism clearly entails token physicalism, but not conversely.

The distinction between token identity theories and type identity theories has not been exploited in the case of behavioristic analyses. Unlike either version of physicalism, behaviorism is generally held as a semantic thesis, hence as a theory about logical relations between types. In the present paper, "physicalism" will mean *type* physicalism. When we talk about states, we will specify whether we mean types or tokens only when it is not clear from the context.

the arguments for this conclusion are widely available in the literature, we shall provide only the briefest review here.[3]

The fundamental argument against behaviorism is simply that what an organism does or is disposed to do at a given time is a very complicated function of its beliefs and desires together with its current sensory inputs and memories. It is thus enormously unlikely that it will prove possible to pair behavioral predicates with psychological predicates in the way that behaviorism requires—namely, that, for each type of psychological state, an organism is in that state if and only if a specified behavioral predicate is true of it. This suggests that behaviorism is overwhelmingly likely to be false simply in virtue of its empirical consequences and independent of its implausibility as a semantic thesis. Behaviorism cannot be true unless mind/behavior correlationism is true, and mind/behavior correlationism is not true.

The argument against physicalism rests upon the empirical likelihood that creatures of different composition and structure, which are in no interesting sense in identical physical states, can nevertheless be in identical psychological states; hence that types of psychological states are not in correspondence with types of physical states. This point has been made persuasively in Putnam's "Psychological Predicates." In essence, it rests on appeals to the following three kinds of empirical considerations.

First, the Lashleyan doctrine of neurological equipotentiality holds that any of a wide variety of psychological functions can be served by any of a wide variety of brain structures. While the generality of this doctrine may be disputed, it does seem clear that the central nervous system is highly labile and that a given type of psychological process is in fact often associated with a variety of distinct neurological structures. (For example, it is a widely known fact that early trauma can lead to the establishment

[3] See Donald Davidson, "Mental Events," in *Fact and Experience*, ed. by Swanson and Foster (Amherst, 1970); Jerry A. Fodor, *Psychological Explanation* (New York, 1968); Hilary Putnam, "Brains and Behavior," in *Analytical Philosophy*, ed. by R. J. Butler (Oxford, 1965); Hilary Putnam, "The Mental Life of Some Machines," in *Modern Materialism: Readings on Mind-Body Identity*, ed. by J. O'Connor (New York, 1966); Hilary Putnam, "Psychological Predicates," in *Art, Mind and Religion*, ed. by Capitan and Merrill (Detroit, 1967).

of linguistic functions in the *right* hemisphere of right-handed subjects.) But physicalism, as we have been construing it, requires that organisms are in type-identical psychological states if and only if they are in type-identical physical states. Hence if equipotentiality is true, physicalism must be false.

The second consideration depends on the assumption that the Darwinian doctrine of convergence applies to the phylogeny of psychology as well as to the phylogeny of morphology and of behavior. It is well known that superficial morphological similarities between organisms may represent no more than parallel evolutionary solutions of the same environmental problem: in particular, that they may be the expression of quite different types of physiological structure. The analogous point about behavioral similarities across species has been widely recognized in the ethological literature: organisms of widely differing phylogeny and morphology may nevertheless come to exhibit superficial behavioral similarities in response to convergent environmental pressures. The present point is that the same considerations may well apply to the phylogeny of the psychology of organisms. Psychological similarities across species may often reflect convergent environmental selection rather than underlying physiological similarities. For example, we have no particular reason to suppose that the physiology of pain in man must have much in common with the physiology of pain in phylogenetically remote species. But if there are organisms whose psychology is homologous to our own but whose physiology is quite different, such organisms provide counterexamples to the psychophysical correlations physicalism requires.

Finally, if we allow the conceptual possibility that psychological predicates could apply to artifacts, then it seems likely that physicalism will prove empirically false. For it seems likely that given any psychophysical correlation which holds for an organism, it is possible to build a machine which is similar to the organism psychologically, but physiologically sufficiently different from the organism that the psychophysical correlation does not hold for the machine.

What these arguments seem to show is that the conditions that behaviorism and physicalism seek to place upon the type identity

of psychological states of organisms are, in a relevant sense, insufficiently abstract. It seems likely that organisms which differ in their behavior or behavioral dispositions can nevertheless be in type-identical psychological states, as can organisms that are in different physical states. (We shall presently discuss a "functionalist" approach to type identity which attempts to set the identity criteria at a level more abstract than physicalism or behaviorism acknowledge.)

Of course, it is *possible* that the type-to-type correspondences required by behaviorism or by physicalism should turn out to obtain. The present point is that even if behavioral or physical states *are* in one-to-one correspondence with psychological states, we have no current evidence that this is so; hence we have no warrant for adopting philosophical theories which *require* that it be so. The paradox about behaviorism and physicalism is that while most of the arguments that have surrounded these doctrines have been narrowly "conceptual," it seems increasingly likely that the decisive arguments against them are empirical.

It is often suggested that one might meet these arguments by supposing that, though neither behavioral nor physical states correspond to psychological states in a one-to-one fashion, they may nevertheless correspond many-to-one. That is, it is supposed that, for each type of psychological state, there is a distinct disjunction of types of behavioral (or physical) states, such that an organism is in the psychological state if and only if it is in one of the disjuncts.

This sort of proposal is, however, shot through with serious difficulties. First, it is less than obvious that there is, in fact, a *distinct* disjunction of behavioral (or physical) states corresponding to each psychological state. For example, there is really no reason to believe that the class of types of behaviors which, in the whole history of the universe, have (or will have) expressed rage for some organism or other, is distinct from the class of types of behaviors which have expressed, say, pain. In considering this possibility, one should bear in mind that practically any behavior might, in the appropriate circumstances, become the conventional expression of practically any psychological state and that a given organism in a given psychological state might exhibit

almost any behavioral disposition depending on its beliefs and preferences. The same kind of point applies, *mutatis mutandis*, against the assumption that there is a distinct disjunction of types of physical states corresponding to each type of psychological state, since it seems plausible that practically any type of physical state could realize practically any type of psychological state in some kind of physical system or other.

But even if there *is* a distinct disjunction of types of behavioral (or physical) states corresponding to each type of psychological state, there is no reason whatever to believe that this correspondence is lawlike; and it is not obvious what philosophical interest would inhere in the discovery of a behavioral (or physical) property which happened, accidentally, to be coextensive with a psychological predicate. Thus, as Davidson has pointed out, on the assumption that psycho-behavioral correlations are not lawlike, even "if we were to find an open sentence couched in behavioral terms and exactly coextensive with some mental predicate, nothing could reasonably persuade us that we had found it" ("Mental Events"). As Davidson has also pointed out, the same remark applies, *mutatis mutandis*, to physical predicates.

Finally, a theory which says that each psychological predicate is coextensive with a distinct disjunction of behavioral (or physical) predicates[4] is incompatible with what we have been assuming is an obvious truth: namely, that a given behavioral state may express (or a given physical state realize) different psychological states at different times. Suppose, for example, that we have a theory which says that the psychological predicate p_1 is coextensive with the disjunctive behavioral predicate α and psychological predicate p_2 is coextensive with the disjunctive behavioral predicate β. Suppose further that S_i designates a type of behavior that has sometimes expressed p_1 but not p_2 and

[4] Not all philosophical behaviorists hold this view; philosophical behaviorism may be broadly characterized as the view that for each psychological predicate there is a behavioral predicate to which it bears a "logical relation." (See Fodor, *op. cit.*) Thus the following view qualifies as behaviorist: all ascriptions of psychological predicates entail ascriptions of behavioral predicates, but not conversely. Though this form of behaviorism is not vulnerable to the present argument, the preceding ones are as effective against it as against biconditional forms of behaviorism.

at other times expressed p_2 but not p_1. Then, S_i will have to be a disjunct of both α and β. But, the disjuncts of α are severally sufficient conditions for p_1 and the disjuncts of β are severally sufficient conditions of p_2 on the assumption that p_1 and α, and p_2 and β, are respectively coextensive. Hence the theory entails that an organism in S_i is in both p_1 and p_2, which is logically incompatible with the claim that S_i expresses p_1 (but not p_2) at some times and p_2 (but not p_1) at others. Of course, one could circumvent this objection by including spatiotemporal designators in the specification of the disjuncts mentioned in α and β. But to do so would be totally to abandon the project of expressing psycho-behavioral (or psychophysical) correlations by lawlike biconditionals.

II

It has recently been proposed that these sorts of difficulties can be circumvented, and an adequate theory of the conditions on type identity of psychological states can be formulated, in the following way. Let us assume that any system P to which psychological predicates can be applied has a description as a probabilistic automaton. (A probabilistic automaton is a generalized Turing machine whose machine table includes instructions associated with finite positive probabilities less than or equal to one. For a brief introduction to the notion of a Turing machine, a machine table, and related notions, see Putnam, "Psychological Predicates.") A *description* of P, in the technical sense intended here, is any true statement to the effect that P possesses distinct states S_1, S_2,. . . S_n which are related to one another and to the outputs and inputs of P by the transition probabilities given in a specified machine table. We will call the states S_1, S_2,. . . S_n specified by the *description* of an organism, the "machine table states of the organism" relative to that *description.*

It is against the background of the assumption that organisms are describable as probabilistic automata that the present theory (hereafter "*FSIT*" for "functional state identity theory") seeks

to specify conditions upon the type identity of psychological states. In particular, *FSIT* claims that for any organism that satisfies psychological predicates at all, there exists a unique best *description* such that each psychological state of the organism is identical with one of its machine table states relative to that description.

Several remarks about *FSIT* are in order. First, there is an obvious generalization of the notion of a probabilistic automaton in which it is thought of as having a separate input tape on which an "oracle" can print symbols during a computation. *FSIT* presupposes an interpretation of this generalization in which sensory transducers take the place of the "oracle" and in which outputs are thought of as instructions to motor transducers. Such an interpretation must be intended if a *description* of an organism is to provide a model of the mental operations of the organism.

Second, we have presented *FSIT* in the usual way as an *identity* theory[5]: in particular, one which claims that each type of psychological state is identical to (a type of) machine table state. Our aim, however, is to sidestep questions about the identity conditions of abstract objects and discuss only a certain class of biconditionals which type-to-type identity statements entail: that is, statements of the form "*O* is in such and such a type of psychological state at time *t* if and only if *O* is in such and such a type of machine table state at time *t*."

Third, it is worth insisting that *FSIT* amounts to more than the claim that every organism has a description as a Turing machine or as a probabilistic automaton. For there are a number of respects in which that claim is trivially true; but its truth in these respects does not entail *FSIT*. For example, if the inputs and outputs of an organism are recursively enumerable (as is the case with any mortal organism after it is dead), then it follows that there exists a Turing machine capable of simulating the organism (that is, a Turing machine which has the same input/output relations). But it does not follow that the organism has a unique best *description* of the sort characterized above.

[5] Cf. Putnam, "Psychological Predicates" and "On Properties," in *Essays in Honor of C. G. Hempel*, ed. by N. Rescher et al. (New York, 1970).

Second, as Putnam has pointed out (in conversation), *everything* is describable as a realization of what one might call the "null" Turing machine: that is, a machine which has only one state and stays in it. (The point is, roughly, that whether a system *P* realizes a Turing machine depends, inter alia, on what counts as a change of state in *P*. If one counts *nothing* as a change of state in *P*, then *P* is a realization of the null Turing machine.) But again, *FSIT* would not be true if the only true *description* of an organism is as a null Turing machine, since *FSIT* requires that the machine table states of an organism correspond one-to-one with its psychological states under its best description.

There are thus two important respects in which *FSIT* involves more than the claim that organisms which satisfy psychological predicates have descriptions. First, *FSIT* claims that such systems have unique best descriptions. Second, *FSIT* claims that the types of machine table states specified by the unique best description of a system are in correspondence with the types of psychological states that the system can be in. It is this second claim of *FSIT* with which we shall be primarily concerned.

FSIT, unlike either behaviorism or physicalism, is not an ontological theory: that is, it is neutral about what token psychological states *are*, in that as far as *FSIT* is concerned, among the systems to which psychological predicates are applicable (and which therefore have *descriptions*) might be included persons, material objects, souls, and so forth. This last point suggests how *FSIT* might meet certain of the kinds of difficulties we raised against physicalism and behaviorism. Just as *FSIT* abstracts from considerations of the ontological status of the systems which have *descriptions*, so too it abstracts from physical differences between systems which have their *descriptions* in common. As Putnam has remarked, "the *same* Turing machine (from the standpoint of the machine table) may be physically realized in a potential infinity of ways" ("The Mental Life of Some Machines," p. 271), and *FSIT* allows us to state type-identity conditions on psychological states which are neutral as between such different realizations.

Similarly, *FSIT* permits us to state such conditions in a way which abstracts from the variety of behavioral consequences

which a psychological state may have. It thereby meets a type of objection which, we supposed above, was fatal to behaviorism.

We remarked that the behaviorist is committed to the view that two organisms are in the same psychological state whenever their behaviors and/or behavioral dispositions are identical; and that this theory is implausible to the extent that the behaviors and the behavioral dispositions of an organism are the effects of *interactions* between its psychological states. But *FSIT* allows us to distinguish between psychological states not only in terms of their behavioral consequences but also in terms of the character of their interconnections. This is because the criterion of identity for machine table states acknowledges *their relations to one another* as well as their relations to inputs and outputs. Thus, *FSIT* can cope with the characteristic indirectness of the relation between psychological states and behavior. Indeed, *FSIT* allows us to see how psychological states which have *no* behavioral expressions might nevertheless be distinct.

Finally, it may be remarked that nothing precludes taking at least some of the transitions specified in a machine table as corresponding to causal relations in the system which the table *describes*. In particular, since *FSIT* is compatible with token physicalism, there is no reason why it should not acknowledge that token psychological states may enter into causal relations. Thus, any advantages which accrue to causal analyses of the psychological states, or of the relations between psychological states and behavior, equally accrue to *FSIT*.[6]

III

In this section we are going to review a series of arguments which provide one degree or another of difficulty for the claim that *FSIT* yields an adequate account of the type-identity conditions for psychological states. It is our view that, taken

[6] Cf. Donald Davidson, "Actions, Reasons and Causes," *Journal of Philosophy*, LX (1963), 685-700.

collectively, these arguments are fairly decisive against the theory of type identity of psychological states that *FSIT* proposes. In the final section we will suggest some reasons why the attempt to provide substantive type-identity conditions on psychological states so often founders.

(1) Any account of type-identity conditions on psychological states that adheres at all closely to our everyday notion of what types of psychological states there are will presumably have to draw a distinction between dispositional states (beliefs, desires, inclinations, and so on) and occurrent states (sensations, thoughts, feelings, and so on). So far as we can see, however, *FSIT* has no plausible way of capturing this distinction short of abandoning its fundamental principle that psychological states correspond one-to-one to machine table states. Suppose, for example, *FSIT* attempts to reconstruct the distinction between occurrents and dispositions by referring to the distinction between the machine table state that an organism is *in* and all the other states specified by its machine table. Thus, one might refine *FSIT* to read: for occurrent states, two organisms are in type-identical psychological states if and only if they are in the same machine table state; and, for each dispositional state, there is a machine table state such that an organism is in the former if and only if its machine table contains the latter.

The following suggests one way of seeing how implausible this proposal is. Every machine table state of an organism is a state in which the organism can be at one time or other. Hence, if the distinction between the machine table state an organism is in and all the other states in its table is the same as the distinction between the occurrent state of an organism and its dispositional states, it follows that every dispositional state of the organism is a possible occurrent state of that organism.

This consequence of *FSIT* is correct for a large number of kinds of psychological dispositions. For example, corresponding to the dispositional predicate "speaks French," there is the occurrent predicate "is speaking French"; corresponding to the dispositional "is greedy" we have the occurrent "is being greedy"; corresponding to the dispositional "can hear sounds above 3,000 Herz" there is "is hearing a sound above 3,000 Herz." And, in

general, for many dispositionals, we have corresponding present progressive forms which denote occurrences.

For many other psychological dispositionals, however, this parallelism fails. For example, we have no "is believing that *P*" corresponding to "believes that *P*"; we have no "is desiring a lollipop" corresponding to "desires a lollipop"; we have no "is preferring *X* to *Y*" corresponding to "prefers *X* to *Y*," and so forth. In short, many dispositional psychological states are *not* possible occurrent psychological states, and for these states *FSIT* offers no obvious model.

It is important to see what this argument does *not* show. According to this argument certain dispositional states cannot correspond to machine table states since all machine table states are possible occurrent states but some dispositional psychological states are not. For such dispositions, there can be no machine table states such that the organism has the former if and only if the latter appears in its *description*. But it is perfectly possible that necessary and sufficient conditions for having such dispositions should be given by reference to some *abstract* property of the organization of machine tables. To take a far-fetched example, given a normal form for descriptions, it might turn out that an organism believes that the sun is 93,000,000 miles from the earth if and only if the first *n* columns in its machine table have some such abstract property as containing only odd integers. Since saying of a machine that the first *n* columns . . . and so forth does not ascribe a machine table state to it, psychological states which are analyzed as corresponding to this sort of property would not thereby be described as possible occurrent states.

To take this line, however, would be to abandon a fundamental claim of *FSIT*. For, while this approach is compatible with the view that two organisms have the same psychology if and only if they have the same machine table, it is *not* compatible with the suggestion that two organisms are in the same (dispositional) psychological state if and only if they have a specified state of their machine tables in common. Hence it is incompatible with the view that psychological states are in one-to-one correspondence with machine table states. Moreover, since we have no way of telling what kinds of abstract properties of machine tables might

turn out to correspond to psychological states, the present difficulty much reduces the possibility of using *FSIT* to delineate substantive type-identity conditions on psychological states. To say that psychological states correspond to some property or other of machine tables is to say something very much weaker than that psychological states correspond to machine table states. This is a kind of point to which we will return later in the discussion.

There is, of course, at least one other way out of the present difficulty for *FSIT*. It might be suggested that we ought to give up the everyday notion that there are some dispositional states which are not possible occurrent states (for example, to acknowledge an occurrent, though perhaps nonconscious, state of believing that *P*). Clearly, the possibility that we might some day have theoretical grounds for acknowledging the existence of such states cannot be precluded a priori. But we have no such grounds *now*, and there does seem to us to be a methodological principle of conservatism to the effect that one should resist models which require empirical or conceptual changes that are not independently motivated.

(2) We suggested that *FSIT* allows us to account for the fact that behavior is characteristically the product of interactions between psychological states, and that the existence of such interactions provides a standing source of difficulty for behaviorist theories in so far as they seek to assign characteristic behaviors *severally* to psychological states. It is empirically immensely likely, however, that there are *two* kinds of behaviorally efficacious interactions between psychological states, and *FSIT* provides for a natural model of only one of them.

On one hand, behavior can be the product of a *series* of psychological states, and the *FSIT* account shows us how this could be true, and how some of the states occurring in such a series may not themselves have behavioral expressions. But, on the other hand, behavior can be the result of interactions between *simultaneous* mental states. For example, prima facie, what an organism does at *t* may be a function of what it is feeling at *t* and what it is thinking at *t*. But *FSIT* provides no conceptual machinery for representing this state of affairs. In effect, *FSIT* can provide for

the representation of sequential interactions between psychological states, but not for simultaneous interactions. Indeed *FSIT* even fails to account for the fact that an organism can be in more than one occurrent psychological state at a time, since a probabilistic automaton can be in only one machine table state at a time. The upshot of this argument seems to be that if probabilistic automata are to be used as models of an organism, the appropriate model will be a set of intercommunicating automata operating in parallel.

It is again important to keep clear on what the argument does not show about *FSIT*. We have read *FSIT* as claiming that the psychological states of an organism are in one-to-one correspondence with the machine table states postulated in its best *description*. The present argument suggests that if this claim is to be accepted, then the best *description* of an organism must not represent it as a single probabilistic automaton. If organisms, but not single probabilistic automata, can be in more than one state at a time, then either an organism is not a single probabilistic automaton, or the psychological states of an organism do not correspond to machine table states of single probabilistic automata. (It should be remarked that there is an algorithm which will construct a single automaton equivalent to any given set of parallel automata. It cannot be the case, however, that a set of parallel automata and the equivalent single automaton *both* provide best *descriptions* of an organism.)

These remarks are of some importance since the kind of psychological theory we get on the assumption that organisms are parallel processors will look quite different from the kind we get on the assumption that they are serial processors. Indeed, while the general characteristics of serial processors are relatively well understood, very little is known about the corresponding characteristics of parallel systems.

On the other hand, this argument does not touch the main claim of *FSIT:* even if organisms are in some sense sets of probabilistic automata, it may turn out that each psychological state of an organism corresponds to a machine table state of one or other of the members of the set. In the following arguments, we will assume the serial model for purposes of simplicity and try

to show that, even on that assumption, psychological states do not correspond to machine table states.

(3) *FSIT* holds that two organisms are in psychological states of the same type if and only if they are in the same machine table state. But machine table states are identical if and only if they are identically related to other machine table states and to states of the input and output mechanisms. In this sense, the criterion for identity of machine table states is "functional equivalence." Thus *FSIT* claims that type identity of psychological states is also a matter of a certain kind of functional equivalence; psychological states are type identical if and only if they share those properties that must be specified to individuate a machine table state.

But it might plausibly be argued that this way of type-identifying psychological states fails to accommodate a feature of at least some such states that is critical for determining their type: namely, their "qualitative" character. It does not, for example, seem entirely unreasonable to suggest that nothing would be a token of the type "pain state" unless it felt like a pain, and that this would be true even if it were connected to all the other psychological states of the organism in whatever ways pains are. It seems to us that the standard verificationist counterarguments against the view that the "inverted spectrum" hypothesis is conceptually coherent are not persuasive. If this is correct, it looks as though the possibility of qualia inversion poses a serious prima-facie argument against functionalist accounts of the criteria for type identity of psychological states.

It should be noticed, however, that the inverted qualia argument is *only* a prima-facie objection against *FSIT*. In particular, it is available to the proponent of functionalist accounts to meet this objection in either of two ways. On the one hand, he might argue that though inverted qualia, *if they occurred*, would provide counterexamples to his theory, as a matter of nomological fact it is impossible that functionally identical psychological states should be qualitatively distinct: in particular, that anything which altered the qualitative characteristics of a psychological state would alter its functional characteristics. This sort of line may strike one as blatant apriorism, but, in the absence of any

relevant empirical data, it might be well to adopt an attitude of wait and see.

There is, moreover, another defense open to the proponent of *FSIT*. He might say that, given two functionally identical psychological states, we would (or perhaps "should") *take* them to be type identical, independent of their qualitative properties: that is, that differences between the qualitative properties of psychological states which do not determine corresponding functional differences are *ipso facto* irrelevant to the goals of theory construction in psychology, and hence should be ignored for purposes of type identification.

To see that this suggestion may be plausible, imagine that it turns out that every person does, in fact, have slightly different qualia (or, better still, grossly different qualia) when in whatever machine table state is alleged to be identical to pain. It seems fairly clear that in this case it might be reasonable to say that the character of an organism's qualia is irrelevant to whether it is in pain or (equivalently) that pains feel quite different to different organisms.

This form of argument may, however, lead to embarrassing consequences. For all that we now know, it may be nomologically possible for two psychological states to be functionally identical (that is, to be identically connected with inputs, outputs, and successor states), even if only one of the states has a qualitative content. In this case, *FSIT* would require us to say that an organism might be in pain even though it is feeling *nothing at all*, and this consequence seems totally unacceptable.

It may be remarked that these "inverted (or absent) qualia" cases in a certain sense pose a deeper problem for *FSIT* than any of the other arguments we shall be discussing. Our other arguments are, by and large, concerned to show that psychological states cannot be functionally defined in a certain way; namely, by being put in correspondence with machine table states. But though they are incompatible with *FSIT*, they are compatible with functionalism in the broad sense of that doctrine which holds that the type-identity conditions for psychological states refer only to their relations to inputs, outputs, and one another. The present consideration, however, might be taken to show

that psychological states cannot be functionally defined *at all* and that they cannot be put into correspondence with *any* properties definable over abstract automata. We will ignore this possibility in what follows, since if psychological states are not functional states at all, the question whether they are machine table states simply does not arise.

(4) We remarked that there are arguments against behaviorism and physicalism which suggest that each proposes constraints upon type-identity conditions on psychological states that are, in one way or another, insufficiently abstract. We will now argue that *FSIT* is susceptible to the same kind of objection.

A machine table specifies a state in terms of a set of instructions which control the behavior of the machine whenever it is in that state. By definition, in the case of a deterministic automaton, such instructions specify, for each state of the machine, an associated output and a successor machine state. Probabilistic automata differ only in that any state may specify a *range* of outputs or of successor states, with an associated probability distribution. In short, two machine table states of a deterministic automaton are distinct if they differ either in their associated outputs or in their associated successor state. Analogously, two machine table states of probabilistic automata differ if they differ in their range of outputs, or in their range of successor states, or in the probability distributions associated with either of these ranges.

If, however, we transfer this convention for distinguishing machine table states to the type identification of psychological states, we get identity conditions which are, as it were, too fine-grained. Thus, for example, if you and I differ *only* in the respect that your most probable response to the pain of stubbing your toe is to say "damn" and mine is to say "darn," it follows that the pain you have when you stub your toe is type-distinct from the pain I have when I stub my toe.

This argument iterates in an embarrassing way. To see this, consider the special case of deterministic automata: x and y are type-distinct machine table states of such an automaton if the immediate successor states of x and y are type-distinct. But the immediate successor states of x and y are type-distinct if *their*

immediate successor states are type-distinct. So x and y are type-distinct if the immediate successors of their immediate successors are type-distinct; and so on. Indeed, on the assumption that there is a computational path from every state to every other, any two automata which have less than all their states in common will have none of their states in common. This argument generalizes to probabilistic automata in an obvious way.

It is again important to see what the argument does *not* show. In particular, it does not show that psychological states cannot be type-identified by reference to some sort of *abstract* properties of machine table states. But, as we remarked in our discussion of Argument 1, to say that psychological states correspond to some or other property definable over machine table states is to say much less about the conditions upon the type identity of psychological states than *FSIT* seeks to do. And the present argument *does* seem to show that the conditions used to type-identify machine table states per se cannot be used to type-identify psychological states. It is presumably this sort of point which Putnam, for example, has in mind when he remarks that "the difficulty of course will be to pass from models of *specific* organisms to a *normal* form for the psychological description of *organisms*" ("Psychological Predicates," p. 43). In short, it may seem at first glance that exploitation of the criteria employed for type-identifying machine table states provides *FSIT* with concepts at precisely the level of abstraction required for type-identifying psychological states. But, in fact, this appears not to be true.

(5) The following argument seems to us to show that the psychological states of organisms cannot be placed in one-to-one correspondence with the machine table states of organisms.

The set of states which constitute the machine table of a probabilistic automaton is, by definition, a list. But the set of mental states of at least some organisms (namely, persons) is, in point of empirical fact, productive. In particular, abstracting from theoretically irrelevant limitations imposed by memory and mortality, there are infinitely many type-distinct, nomologically possible psychological states of any given person. The simplest demonstration that this is true is that, on the assumption that there are infinitely many non-equivalent de-

clarative sentences, one can generate definite descriptions of such states by replacing *S* with sentences in the schemata *A*:

A: "the belief (thought, desire, hope, and so forth) that *S*"

In short, while the set of machine table states of a Turing machine can, by definition, be exhaustively specified by listing them, the set of mental states of a person can at best be specified by finite axiomatization.

It might be maintained against this argument that not more than a finite subset of the definite descriptions generable by substitution in *A* do in fact designate nomologically possible beliefs (desires, hopes, or whatever) and that this is true *not* because of theoretically uninteresting limitations imposed by memory and mortality, but rather because of the existence of psychological laws that limit the set of believable (and so forth) propositions to a finite set. To take a farfetched example, it might be held that if you eliminate all such perhaps unbelievable propositions as "2 + 2 = 17," "2 + 2 = 147," and so forth, the residuum is a finite set.

There is no reason at all to believe that this is true, however, and there are some very persuasive reasons for believing that it is not. For example, the infinite set of descriptions whose members are "the belief that 1 + 1 = 2," "the belief that 2 + 2 = 4," "the belief that 3 + 3 = 6," and so forth would appear to designate a set of possible beliefs of an organism ideally free from limitations on memory; to put it the other way around, the fact that there are arithmetical statements that it is nomologically impossible for any person to believe is a consequence of the character of people's memory, not a consequence of the character of their mental representation of arithmetic.

It should be emphasized, again, that this is intended to be an empirical claim, albeit an overwhelmingly plausible one. It is possible to imagine a creature ideally free from memory limitations whose mental representation of arithmetic nevertheless specifies only a finite set of possible arithmetic beliefs. The present point is that it is vastly unlikely that we are such creatures.

Once again it is important to see what the argument does *not* show. Let us distinguish between the *machine table states* of an

automaton, and the *computational states* of an automaton. By the former, we will mean what we have been meaning all along: states specified by columns in its machine table. By the latter we mean any state of the machine which is characterizable in terms of its inputs, outputs, and/or machine table states. In this usage, the predicates "has just run through a computation involving three hundred seventy-two machine table states," or "has just proved Fermat's last theorem," or "has just typed the *i*th symbol in its output vocabulary" all designate possible computational states of machines.

Now, what the present argument seems to show is that the psychological states of an organism cannot be put into correspondence with the machine table states of an automaton. What it of course does *not* show is that the psychological states of an organism cannot be put into correspondence with the *computational* states of an automaton. Indeed, a sufficient condition for the existence of the latter correspondence is that the psychological states of an organism should be countable.[7]

(6) We have argued that since the set of machine table states of an automaton is not productive, it cannot be put into correspondence with the psychological states of an organism. We will now argue that even if such a correspondence could be effected, it would necessarily fail to represent essential properties of psychological states. It seems fairly clear that there are structural

[7] The claim that organisms are probabilistic automata might be interestingly true even if *FSIT* is false; that is, even if psychological states do not correspond to machine table states. For example, it might turn out that some subset of the psychological states of an organism correspond to a set of machine table states by which the rest of its psychology is determined. Or it might turn out that what corresponds to each machine table state is a *conjunction* of psychological states . . . , etc. Indeed, though the claim that any organism can be modeled by a probabilistic automaton is not interesting, the claim that for each organism there is a probabilistic automaton which is its *unique best* model *is* interesting. And this latter claim neither entails *FSIT* nor is it by any means obviously true.

In short, there are many ways in which it could turn out that organisms are automata in some sense *more* interesting than the sense in which everything is an automaton under some description. Our present point is that such eventualities, while they would be important, would not provide general conditions upon the type identification of psychological states in the way that *FSIT* attempts to do.

similarities among at least some psychological states, and that a successful theory of such states must represent and exploit such similarities. For example, there is clearly some theoretically relevant relation between the psychological state that someone is in if he believes that P and the psychological state that someone is in if he believes that P & Q. The present point is simply that representing the psychological states as a list (for example, as a list of machine table states) fails to represent this kind of structural relation. What needs to be said is that believing that P is somehow[8] a constituent of believing that P & Q; but the machine table state model has no conceptual resources for saying that. In particular, the notion "is a constituent of" is not defined for machine table states.

It might be replied that this sort of argument is not strictly relevant to the claims of *FSIT*: for it is surely possible, in principle, that there should be a one-to-one correspondence between machine table states and psychological states, even though the vocabulary appropriate to the individuation of the former does not capture the structural relations among the latter.

This reply, however, misses the point. To see this, consider the case with sentences. The reason there are structural parallelisms among sentences is that sentences are constructed from a fixed set of vocabulary items by the iterated application of a fixed set of rules, and the theoretical vocabulary required to analyze the ways in which sentences are structurally similar is precisely the vocabulary required to specify the domain of those rules. In particular, structurally similar sentences share either lexical items or paths in their derivations, or both. Thus one explains structural similarities between sentences in the same way that one explains their productivity: namely, by describing them as a generated set rather than a list.

Our point is that the same considerations apply to the set of psychological states of an organism. Almost certainly, they too are, or at least include, a generated set, and their structural

[8] Very much "somehow." Obviously, believing p is not a constituent of believing $p \vee q$ in the same way that believing p is a constituent of believing p & q. Equally obviously, there is some relation between believing p and believing $p \vee q$, and a theory of belief will have to say what that relation is.

similarities correspond, at least in part, to similarities in their derivation; that is, with psychological states as with sentences, the fact that they are productive and the fact that they exhibit internal structure are two aspects of the same phenomenon. If this is true, then a theory which fails to capture the structural relations within and among psychological states is overwhelmingly unlikely to arrive at a *description* adequate for the purposes of theoretical psychology.

This argument, like 5, thus leads to the conclusion that, if we wish to think of the psychology of organisms as represented by automata, then the psychological states of organisms seem to be analogous to the computational states of an automaton rather than to its machine table states.

IV

We have been considering theories in the philosophy of mind which can be construed as attempts to place substantive conditions upon type identity of psychological states. We have argued that none of the major theories currently in the field appear to succeed in this enterprise. It might, therefore, be advisable to reconsider the whole undertaking.

Suppose someone wanted to know what the criteria for type identity of fundamental physical entities are. Perhaps the best one could do by way of answering is to say that two such entities are type-identical if they do not differ with respect to any fundamental physical magnitudes. Thus, as far as we know, the conditions upon type identification of elementary physical particles do not refer to their distance from the North Pole, but do refer to their charge. But notice that this is simply a consequence of the fact that there are no fundamental physical laws which operate on entities as a function of their distance from the North Pole, and there *are* fundamental physical laws which operate on entities as a function of their charge.

One might put it that the basic condition upon type identity in science is that it makes possible the articulation of the domain of laws. This principle holds at every level of scientific description.

Thus what is *relevant* to the question whether two entities at a level will be type-distinct is the character of the laws which operate upon entities at that level. But if this is the general case, then it looks as though substantive conditions upon type identity of psychological states will be imposed by reference to the psychological (and perhaps neurological) laws which operate upon those states and in no other way.

In the light of these remarks, we can get a clearer view of what has gone wrong with the kinds of philosophical theories we have been rejecting. For example, one can think of behaviorism as involving an attempt to type-identify psychological states just by reference to whatever laws determine their *behavioral* effects. But this would seem, even prima facie, to be a mistake, since there must be laws which govern the interaction of psychological states and there is no reason to believe (and much reason not to believe) that psychological states which behave uniformly vis-à-vis laws of the first kind characteristically behave uniformly vis-à-vis laws of the second kind.

Analogously, what has gone wrong in the case of physicalism is the assumption that psychological states that are distinct in their behavior vis-à-vis neurological laws are *ipso facto* distinct in their behavior vis-à-vis psychological laws. But, in all probability, distinct neurological states can be functionally identical. That is, satisfaction of the criteria for type-distinctness of neurological states probably does not guarantee satisfaction of the criteria for type-distinctness of psychological states or vice versa.

In short, the fundamental problem with behaviorism and physicalism is that type identity is being determined relative to, at best, a subset of the laws which must be presumed to operate upon psychological states. The only justification for this restriction seems to lie in the reductionist biases of these positions. Once the reductionism has been questioned, we can see that the nomological demands upon type identification for psychological states are likely to be extremely complicated and various. Even what little we already know about psychological laws makes it look implausible that they will acknowledge type boundaries between psychological states at the places where physicalists or behaviorists want to draw them.

The basic failure of *FSIT* is in certain respects analogous to that of behaviorism and physicalism. Of course, *FSIT* is not reductionist even in spirit, and in so far as it is a species of functionalism it does invite us to type-identify psychological states by reference to their nomological connections with sensory inputs, behavioral outputs, and with one another. But *FSIT* seeks to impose a further constraint on type identity: namely, that the psychological states of an organism can be placed in correspondence with (indeed, identified with) the machine table states specified by the best *description* of the organism. We have argued that this is in fact a substantive constraint, and one which cannot be satisfied.

What seems to be left of *FSIT* is this. It may be both true and important that organisms are probabilistic automata. But even if it is true and important, the fact that organisms are probabilistic automata seems to have very little or nothing to do with the conditions on type identity of their psychological states.

N. J. BLOCK and J. A. FODOR

Massachusetts Institute of Technology