MIND DESIGN

Philosophy
Psychology
Artificial Intelligence

edited by

JOHN HAUGELAND

A Bradford Book

The MIT Press
Cambridge, Massachusetts
London, England

1975, and published in *Communications of the Association for Computing Machinery*, 19 (March 1976) 113–126. It is reprinted by permission of the ACM.

2. "Complexity and the Study of Artificial and Human Intelligence," by Zenon W. Pylyshyn, was presented (in an earlier version) at a conference on Objectives and Methodologies for Artificial Intelligence, in Canberra, Australia (May 1974), and first published in *Philosophical Perspectives in Artificial Intelligence*, edited by Martin Ringle (Atlantic Highlands, N.J.: Humanities Press, 1979). It is reprinted here (with modest revisions) by permission of Humanities Press.
3. "A Framework for Representing Knowledge," by Marvin Minsky, was originally published as Memo 306 of the Artificial Intelligence Laboratory at MIT. Excerpts were reprinted in *The Psychology of Computer Vision*, edited by Patrick H. Winston (New York: McGraw Hill, 1975); and other excerpts were reprinted in the Proceedings of the 1975 TINLAP Conference, in Cambridge, Massachusetts. Still other excerpts are reprinted here by permission of Professor Minsky.
4. "Artificial Intelligence—A Personal View," by David Marr, was first published in *Artificial Intelligence*, 9 (1977) 37–48. It is reprinted here by permission of North-Holland Publishing Company.
5. "Artificial Intelligence Meets Natural Stupidity," by Drew McDermott, was first published in the *SIGART Newsletter* (of the Special Interest Group on Artificial Intelligence, of the Association for Computing Machinery), No. 57 (April 1976). It is reprinted here by permission of Professor McDermott.
6. "From Micro-Worlds to Knowledge Representation: AI at an Impasse," by Hubert L. Dreyfus, is excerpted (with minor revisions) from the Introduction to the second edition of his *What Computers Can't Do* (New York: Harper and Row, 1979). It is reprinted here by permission of Harper and Row.
7. "Reductionism and the Nature of Psychology," by Hilary Putnam, was first published in *Cognition*, 2 (1973) 131–146. It is reprinted here (somewhat abridged) by courtesy of Elsevier Sequoia, Lausanne.

8. "Intentional Systems," by Daniel C. Dennett, was first published in *The Journal of Philosophy*, 68 (1971) 87–106; it has been reprinted in Professor Dennett's *Brainstorms* (Montgomery, Vermont: Bradford Books, 1978). It is reprinted here by permission of *The Journal of Philosophy*.
9. "The Nature and Plausibility of Cognitivism," by John Haugeland, was first published in *The Behavioral and Brain Sciences*, 1 (1978) 215–226. Copyright © 1978 Cambridge University Press. Reprinted (with minor revisions) by permission of the publisher.
10. "Minds, Brains, and Programs," by John R. Searle, was first published in *The Behavioral and Brain Sciences*, 3 (1980), 417–424. Copyright © 1980 Cambridge University Press. Reprinted by permission of the publisher.
11. "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology," by Jerry A. Fodor, was first published in *The Behavioral and Brain Sciences*, 3 (1980), 63–73. Copyright © 1980 Cambridge University Press. Reprinted by permission of the publisher.
12. "The Material Mind," by Donald Davidson, was first published in *Logic, Methodology and Philosophy of Science IV*, edited by Pat Suppes, et al. (Amsterdam: North-Holland, 1973). It is reprinted by permission of North-Holland Publishing Company.

attempts to overcome my ignorance of artificial intelligence. I would especially like to thank Ned Block, Hubert Dreyfus, John Haugeland, Roger Schank, Robert Wilensky, and Terry Winograd.

11

Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology

JERRY A. FODOR

> . . . to form the idea of an object and to form an idea simply is the same thing; the reference of the idea to an object being an extraneous denomination, of which in itself it bears no mark or character.
>
> —Hume (1888), p. 20

THE PAPER distinguishes two doctrines, both of which inform theory construction in much of modern cognitive psychology: the representational theory of mind (according to which propositional attitudes are relations that organisms bear to mental representations) and the computational theory of mind (according to which mental processes have access only to formal (nonsemantic) properties of the mental representations over which they are defined.

It is argued that the acceptance of some such formality condition is warranted, at least for that part of psychology which concerns itself with the mental causation of behavior. The paper closes with a discussion of the prospects for a "naturalistic" psychology: one which defines its generalizations over relations between mental representations and their environmental causes. Two related arguments are proposed, both leading to the conclusion that no such research strategy is likely to prove fruitful.

Your standard contemporary cognitive psychologist—your thoroughly modern mentalist—is disposed to reason as follows. To think (e.g.,) that Marvin is melancholy is to represent Marvin in a certain way; viz. as being melancholy (and not, for example,

as being maudlin, morose, moody, or merely moping and dyspeptic). But surely we cannot represent Marvin as being melancholy except as we are in some or other relation to a representation of Marvin; and not just to *any* representation of Marvin, but, in particular, to a representation the content of which is *that* Marvin is melancholy; a representation which, as it were, expresses the proposition that Marvin is melancholy. So, a fortiori, at least some mental states/processes are or involve at least some relations to at least some representations. Perhaps, then, this is the *typical* feature of such mental states/processes as cognitive psychology studies; perhaps all such states can be viewed as relations to representations and all such processes as operations defined on representations.

This is, prima facie, an appealing proposal, since it gives the psychologist two degrees of freedom to play with and they seem, intuitively, to be the right two. On the one hand, mental states are distinguished by the *content* of the associated representations, and we therefore can allow for the difference between thinking that Marvin is melancholy and thinking that Sam is (or that Albert isn't, or that it sometimes snows in Cincinnati); and, on the other hand, mental states are distinguished by the *relation* that the subject bears to the associated representation (so we can allow for the difference between thinking, hoping, supposing, doubting and pretending that Marvin is melancholy). It's hard to believe that a serious psychology could make do with fewer (or less refined) distinctions than these, and it's hard to believe that a psychology that makes these distinctions could avoid taking the notion of mental representation seriously. Moreover, the burden of argument is clearly upon anyone who claims that we need *more* degrees of freedom than just these two: the least hypothesis that is remotely plausible is that a mental state is (type) individuated by specifying a relation and a representation such that the subject bears the one to the other.[1]

1. I shall speak of 'type identity' (distinctness) of mental states to pick out the sense of 'same mental state' in which, for example, John and Mary are in the same mental state if both believe that water flows. Correspondingly, I shall use the notion of 'token identity' (distinctness) of mental state to pick out the sense of 'same mental state' in which it's necessary that if x and y are in the same mental state, then $x = y$.

I'll say that any psychology that takes this line is a version of the REPRESENTATIONAL THEORY OF THE MIND. I think that it's reasonable to adopt some such theory as a sort of working hypothesis, if only because there aren't any alternatives which seem to be even *remotely* plausible and because empirical research carried out within this framework has, thus far, proved interesting and fruitful.[2] However, my present concern is neither to attack nor to defend this view, but rather to distinguish it from something other—and stronger—that modern cognitive psychologists *also* hold. I shall put this stronger doctrine as the view that mental states and processes are COMPUTATIONAL. Much of what is characteristic of cognitive psychology is a consequence of adherence to this stronger view. What I want to do in this paper is to say something about what this stronger view is, something about why I think it's plausible, and, most of all, something about the ways in which it shapes the cognitive psychology we have.

I take it that computational processes are both *symbolic* and *formal*. They are symbolic because they are defined over representations, and they are formal because they apply to representations in virtue of (roughly) the *syntax* of the representations. It's the second of these conditions that makes the claim that mental processes are computational stronger than the representational theory of the mind. Most of this paper will be a meditation upon the consequences of assuming that mental processes are formal processes.

I'd better cash the parenthetical 'roughly'. To say that an operation is formal isn't the same as saying that it is syntactic since we could have formal processes defined over representations which don't, in any obvious sense *have* a syntax. Rotating an image would be a timely example. What makes syntactic operations a species of formal operations is that being syntactic is a way of *not* being semantic. Formal operations are the ones that are specified without reference to such semantic properties of representations as, for example, truth, reference and meaning. Since we don't know how to complete this list (since, that is, we don't know what semantic properties there are), I see no responsible way of saying what, in general, formality amounts

2. For extensive discussion, see Fodor (1975, 1978b).

to. The notion of formality will thus have to remain intuitive and metaphoric, at least for present purposes: formal operations apply in terms of the, as it were, shapes of the objects in their domains.[3]

To require that mental processes be computational (viz. formal-syntactic) is thus to require something not very clear. Still, the requirement has some clear consequences, and they are striking and tendentious. Consider that we started by assuming that the *content* of representations is a (type) individuating feature of mental states. So far as the *representational* theory of the mind is concerned, it's possibly the *only* thing that distinguishes Peter's thought that Sam is silly from his thought that Sally is depressed. But, now, if the *computational* theory of the mind is true (and if, as we may assume, content is a semantic notion par excellence) it follows that content alone cannot distinguish thoughts. More exactly, the computational theory of the mind requires that two thoughts can be distinct in content only if they can be identified with relations to formally distinct representations. More generally: fix the subject and the relation, and then mental states can be (type) distinct only if the representations which constitute their objects are formally distinct.

Again, consider that accepting a formality condition upon mental states implies a drastic narrowing of the ordinary ontology of the mental; all sorts of states which look, prima facie, to be mental in good standing are going to turn out to be none of the psychologist's business if the formality condition is endorsed. This point is one that philosophers have made in a number of contexts, and usually in a deprecating tone of voice. Take, for example, knowing that such-and-such, and assume that you can't know what's not the case. Since, on that assumption, knowledge is involved with truth, and since truth is a semantic notion, it's going to follow that there can't be a psychology of *knowledge* (even if it is consonant with the formality condition to hope for a psychology of *belief*). Similarly, it's a way of making a point

3. This is *not*, notice, the same as saying 'formal operations are the ones that apply mechanically'; in this latter sense, *formality* means something like *explicitness*. There's no particular reason for using 'formal' to mean both 'syntactic' and 'explicit', though the ambiguity abounds in the literature.

of Ryle's to say that, strictly speaking, there can't be a psychology of perception if the formality condition is to be complied with. Seeing is an achievement; you can't see what's not there. From the point of view of the representational theory of the mind, this means that seeing involves relations between mental representations *and their referents;* hence, semantic relations within the meaning of the act.

I hope that such examples suggest (what, in fact, I think is true) that even if the formality condition isn't very clear, it is quite certainly very strong. In fact, I think it's not all *that* anachronistic to see it as the central issue which divides the two main traditions in the history of psychology: 'Rational psychology' on the one hand, and 'Naturalism' on the other. Since this is a mildly eccentric way of cutting the pie, I'm going to permit myself a semihistorical excursus before returning to the main business of the paper.

Descartes argued that there is an important sense in which how the world is makes no difference to one's mental states. Here is a well known passage from the first *Meditation:*

> At this moment it does indeed seem to me that it is with eyes awake that I am looking at this paper; that this head which I move is not asleep, that it is deliberately and of set purpose that I extend my hand and perceive it . . . But in thinking over this I remind myself that on many occasions I have been deceived by similar illusions, and in dwelling on this reflection I see so manifestly that there are no certain indications by which we may clearly distinguish wakefulness from sleep that I am lost in astonishment. And my astonishment is such that it is almost capable of persuading me that I now dream. (1967; p. 146)

At least three sorts of reactions to this kind of argument are distinguishable in the philosophical literature. First, there's a long tradition, including both Rationalists and Empiricists, which takes it as axiomatic that one's experiences (and, a fortiori, one's beliefs) might have been just as they are even if the world had been quite different from the way that it is. See, for example, the passage from Hume which serves as an epigraph to this paper. Second, there's a vaguely Wittgensteinian mood in which one argues that it's just *false* that one's mental states might have been what they

are had the world been relevantly different. For example, if there had been a dagger there, Macbeth would have been *seeing*, not just hallucinating. And what could be more different than that? If the Cartesian feels that this reply misses the point, he is at least under an obligation to say precisely which point it misses; in precisely *which* respects the way the world is is irrelevant to the character of one's beliefs, experiences, etc. Finally there's a tradition which argues that—epistemology to one side—it is at best a strategic mistake to attempt to develop a psychology which individuates mental states without reference to their environmental causes and effects (e.g., which counts the state that Macbeth *was* in as type-identical to the state he would have been in had the dagger been supplied.) I have in mind the tradition which includes the American Naturalists (notably Pierce and Dewey), all the learning theorists, and such contemporary representatives as Quine in philosophy and Gibson in psychology. The recurrent theme here is that psychology is a branch of biology, hence that one must view the organism as embedded in a physical environment. The psychologist's job is to trace those organism/environment interactions which constitute its behavior. A passage from William James (1890) will serve to give the feel of the thing:

> On the whole, few recent formulas have done more service of a rough sort in psychology than the Spencerian one that the essence of mental life and of bodily life are one, namely, 'the adjustment of inner to outer relations.' Such a formula is vagueness incarnate; but because it takes into account the fact that minds inhabit environments which act on them and on which they in turn react; because, in short, it takes mind in the midst of all its concrete relations, it is immensely more fertile than the old-fashioned 'rational psychology' which treated the soul as a detached existent, sufficient unto itself, and assumed to consider only its nature and its properties. (p. 6)

A number of adventitious intrusions have served to muddy the issues in this long-standing dispute. On the one hand, it may well be that Descartes was relying on a specifically introspectionist construal of the claim that the individuation of mental states is independent of their environmental causes. That is, Descartes' point may have been that (a) mental states are (type) identical if

and only if they are introspectively indistinguishable, and (b) introspection cannot distinguish (e.g.,) perception from hallucination, or knowledge from belief. On the other hand, the naturalist, in point of historical fact, is often a behaviorist as well. He wants to argue not only that mental states are individuated by reference to organism/environment relations, but also that such relations constitute the mental. In the context of the present discussion, he is arguing for the abandonment not just of the formality condition, but of the notion of mental representation as well.

If, however, we take the computational theory of the mind as what's central to the issue, we can reconstruct the debate between rational psychologists and naturalists in a way that does justice to both their points; in particular, in a way which frees the discussion from involvement with introspectionism on the one side and behaviorism on the other.

Insofar as we think of mental processes as computational (hence as formal operations defined on representations) it will be natural to take the mind to be, *inter alia*, a kind of computer. That is, we will think of the mind as carrying out whatever symbol manipulations are constitutive of the hypothesized computational processes. To a first approximation, we may thus construe mental operations as pretty directly analogous to those of a Turing machine. There is, for example, a working memory (corresponding to a tape) and there are capacities for scanning and altering the contents of the memory (corresponding to the operations of reading and writing on the tape). If we want to extend the computational metaphor by providing access to information about the environment, we can think of the computer as having access to "oracles" which serve, on occasion, to enter information in the memory. On the intended interpretation of this model, these oracles are analogs to the senses. In particular, they are assumed to be transducers, in that what they write on the tape is determined solely by the ambient environmental energies that impinge upon them. (For elaboration of this sort of account, see Putnam, 1960; it is, of course, widely familiar from discussions in artificial intelligence.)

I'm not endorsing this model, but simply presenting it as a natural extension of the computational picture of the mind. Its present interest is that we can use it to see how the formality condition connects with the Cartesian claim that the character

of mental processes is somehow independent of their environmental causes and effects. The point is that, so long as we are thinking of mental processes as purely computational, the bearing of environmental information upon such processes is exhausted by the formal character of whatever the oracles write on the tape. In particular, it doesn't matter to such processes whether what the oracles write is *true*; whether, for example, they really are transducers faithfully mirroring the state of the environment, or merely the output end of a typewriter manipulated by a Cartesian demon bent on deceiving the machine. I'm saying, in effect, that the formality condition, viewed in this context, is tantamount to a sort of methodological solipsism. If mental processes are formal, they have access only to the formal properties of such representations of the environment as the senses provide. Hence, they have no access to the *semantic* properties of such representations, including the property of being true, of having referents, or, indeed, the property of being representations *of the environment.*

That some such methodological solipsism really is implicated in much current psychological practice is best seen by examining what researchers actually do. Consider, for example, the well-known work of Professor Terry Winograd. Winograd was primarily interested in the computer simulation of certain processes involved in the handling of verbal information; asking and answering questions, drawing inferences, following instructions and the like.The form of his theory was a program for a computer which 'lives in' and operates upon a simple world of block-like geometric objects. (Cf. Winograd, 1971) Many of the capacities that the device exercises vis-à-vis its environment seem impressively intelligent. It can arrange the blocks to order, it can issue 'perceptual' reports of the present state of its environment and 'memory' reports of its past states, it can devise simple plans for achieving desired environment configurations, and it can discuss its undertakings (more or less in English) with whoever is running the program.

The interesting point for our purposes, however, is that the machine environment which is the nominal object of these actions and conversations actually isn't there. What actually happens is that the programmer so arranges the memory states of the machine that the available data are whatever they would be *if* there were objects for the machine to perceive and manipulanda for it to operate upon. In effect, the machine lives in an entirely notional world; all its beliefs are false. Of course, it doesn't matter to the machine that its beliefs are false since falsity is a semantic property and, qua computer, the device satisfies the formality condition; viz. it has access only to formal (non-semantic) properties of the representations that it manipulates. In effect, the device is in precisely the situation that Descartes dreads; it's a mere computer which dreams that it's a robot.

I hope that this discussion suggests how acceptance of the computational theory of the mind leads to a sort of methodological solipsism as a part of the research strategy of contemporary cognitive psychology. In particular, I hope it's clear how you get that consequence from the formality condition alone, without so much as raising the introspection issue. I stress this point because it seems to me that there has been considerable confusion about it among the psychologists themselves. People who do machine simulation, in particular, very often advertise themselves as working on the question how thought (or language) is related to the world. My present point is that, whatever else they're doing, they certainly aren't doing *that.* The very assumption that defines their field—viz. that they study mental processes *qua* formal operations on symbols—guarantees that their studies won't answer the question how the symbols so manipulated are semantically interpreted. You can, for example, build a machine that answers baseball questions in the sense that (e.g.) if you type in "Who had the most wins by a National League pitcher since Dizzy Dean?" it will type out "Robin Roberts, who won 28." But you delude yourself if you think that a machine which in this sense answers baseball questions is thereby answering questions *about* baseball (or that the machine has somehow referred to Robin Roberts). If the *programmer* chooses to interpret the machine inscription "Robin Roberts won 28" as a statement about Robin Roberts (e.g., as the statement that he won 28), that's all well and good, but it's no business of the machine's. The machine has no access to that interpretation, and its computations are in no way affected by it. The machine doesn't know what it's talking about, and it doesn't care; *about* is a semantic relation.[4]

4. Some fairly deep methodological issues in AI are involved here. See Fodor (1978a), where this surface is lightly scratched.

This brings us to a point where, having done some sort of justice to the Cartesian's insight, we can also do some sort of justice to the naturalist's. For, after all, mental processes are supposed to be operations on representations, and it is in the nature of representations to represent. We have seen that a psychology which embraces the formality condition is thereby debarred from raising questions about the semantic properties of mental representations; yet surely such questions ought *somewhere* to be raised. The computer which prints out 'RR won 28" is not thereby referring to RR. But, surely, when I think *RR won 28,* I *am* thinking about RR, and if not in virtue of having performed some formal operations on some representations, then presumably in virtue of something else. It's perhaps borrowing the least tendentious fragment of causal theories of reference to assume that what fixes the interpretation of my mental representations of RR is something about the way that he and I are embedded in the world; perhaps not a causal chain stretching between us, but anyhow *some* facts about how he and I are causally situated; *Dasein,* as you might say. Only a *naturalistic* psychology will do to specify these facts, because here we are explicitly in the realm of organism/environment transactions.

We are on the verge of a bland and ecumenical conclusion: that there is room both for a computational psychology—viewed as a theory of formal processes defined over mental representations—*and* a naturalistic psychology, viewed as a theory of the (presumably causal) relations between representations and the world which fix their semantic interpretations of the former. I think that, in principle, this is the right way to look at things. In practice, however, I think that it's misleading. So far as I can see, it's overwhelmingly likely that computational psychology is the only one that we are going to get. I want to argue for this conclusion in two steps. First, I'll argue for what I've till now only assumed: that we must *at least* have a psychology which accepts the formality condition. Then I'll argue that there's good reason to suppose that that's the most that we can have; that a naturalistic psychology isn't a practical possibility and isn't likely to become one.

The first move, then, is to give reasons for believing that at least *some* part of psychology should honor the formality condition. Here too the argument proceeds in two steps. I'll argue first that

it is typically under an *opaque* construal that attributions of propositional attitudes to organisms enter into explanations of their behavior; and second that the formality condition is intimately involved with the explanation of propositional attitudes so construed: roughly, that it's reasonable to believe that we can get such explanations only within computational theories. *Caveat emptor:* the arguments under review are, in large part, nondemonstrative. In particular, they will assume the perfectibility in principle of the kinds of psychological theories now being developed, and it is entirely possible that this is an assumption contrary to fact.

Thesis: when we articulate the generalizations in virtue of which behavior is contingent upon mental states, it is typically an opaque construal of the mental state attributions that does the work; for example, it's a construal under which believing that *a is F* is logically independent from believing that *b is F,* even in the case where a = b. It will be convenient to speak not only of opaque construals of propositional attitude ascriptions, but also of *opaque taxonomies* of mental state types; e.g. of taxonomies which, *inter alia,* count the belief that the Morning Star rises in the East as type distinct from the belief that the Evening Star does. (Correspondingly, *transparent* taxonomies are such as, *inter alia,* would count these beliefs as type identical.) So, the claim is that mental states are typically opaquely taxonomized for purposes of psychological theory.[5]

The point doesn't depend upon the examples, so I'll stick to the most informal sorts of cases. Suppose I know that John wants to meet the girl who lives next door; and suppose I know that this is true when 'wants to' is construed opaquely. Then, given even

5. I'm told by some of my friends that this paragraph could be read as suggesting that there are *two kinds* of beliefs: opaque ones and transparent ones. That is not, of course, the way that it is intended to be read. The idea is rather that there are two kinds of conditions that we can place on determinations that a pair of belief tokens count as tokens of the same belief type. According to one set of conditions (corresponding to transparent taxonomy), a belief that the Morning Star is such and such counts as the same belief as a belief that the Evening Star is such and such; whereas, according to the other set of conditions (corresponding to opaque taxonomy), it does not.

rough-and-ready generalizations about how people's behaviors are contingent upon their utilities, I can make some reasonable predictions (/guesses) about what John is likely to do: he's likely to say (viz. utter), "I want to meet the girl who lives next door". He's likely to call upon his neighbor. He's likely (at a minimum, and all things being equal) to exhibit next-door-directed behavior. None of this is frightfully exciting, but it's all I need for present purposes, and what more would you expect from folk psychology?

On the other hand, suppose that all I know is that John wants to meet the girl next door where 'wants to' is construed transparently. I.e., all I know is that it's true of the girl next door that John wants to meet her. Then there is little or nothing that I can predict about how John is likely to proceed. And this is *not* just because rough and ready psychological generalizations want *ceteris paribus* clauses to fill them in; it's also for the deeper reason that I can't infer from what I know about John to any relevant description of the mental causes of his behavior. For example, I have no reason to predict that John will say such things as "I want to meet the girl who lives next door" since, let John be as cooperative and as truthful as you like, and let him be utterly a native speaker, still, he *may believe* that the girl he wants to meet languishes in Latvia. In which case, "I want to meet the girl who lives next door" is the last thing it will occur to him to say. (The contestant wants to say 'suspender', for 'suspender' is the magic word. Consider what we can predict about his probable verbal behavior if we take this (a) opaquely and (b) transparently. And, of course, the same sorts of points apply, *mutatis mutandis*, to the prediction of *non*verbal behavior).

Ontologically, transparent readings are stronger than opaque ones; for example, the former license existential inferences which the latter do not. But psychologically, opaque readings are stronger than transparent ones; they tell us more about the character of the mental causes of behavior. The representational theory of mind offers an explanation of this anomaly. Opaque ascriptions are true in virtue of the way that the agent represents the objects of his wants (intentions, beliefs, etc.) *to himself*. And, by assumption, such representations function in the causation of the behaviors that the agent produces. So, for example, to say that it's true *opaquely* that Oedipus did such-and-such because he wanted

to marry Jocasta, is to say something like (though not, perhaps, *very* like; see Fodor, 1978b): "Oedipus said to himself, 'I want to marry Jocasta', and his so saying was among the causes of his behavior". Whereas to say (only) that it's true transparently that O. wanted to marry J. is to say no more than that among the causes of his behavior was O's saying to himself 'I want to marry . . .' where the blank was filled by *some* expression that denotes J.[6] But now, what O. *does*, how he in the proprietary sense behaves, will depend on which description he (literally) had in mind.[7] If it's 'Jocasta', courtship behavior follows *ceteris paribus*. Whereas, if it's 'my Mum', we have the situation towards the end of the play and Oedipus at Colonus eventually ensues.

I dearly wish that I could leave this topic here, because it would be very convenient to be able to say, without qualification, what I strongly implied above: the opaque readings of propositional attitude ascriptions tell us how people represent the objects of

6. I'm leaving it open that it may be to say still less than this (e.g., because of problems about reference under false descriptions). For purposes of the present discussion, I don't need to run a line on the truth conditions for transparent propositional attitude ascriptions. Thank Heaven, since I do not have one.

7. It's worth emphasizing that the sense of 'behavior' *is* proprietary, and that that's pretty much what you would expect. Not every true description of an act can be such that a theory of the mental causation of behavior will explain the act under that description. (In being rude to Darcy, Elizabeth is insulting the man whom she will eventually marry. A theory of the mental causation of her behavior might have access to the former description, but not, surely, to the latter.)

Many philosophers—especially since Wittgenstein—have emphasized the ways in which the description of behavior may depend upon its context, and it is a frequent charge against modern versions of Rational psychology that they typically ignore such characterizations. So they do, but so what? You can't have explanations of everything under every description, and it's a question for empirical determination which descriptions of behavior reveal its systematicity vis-à-vis its causes. The Rational psychologist is prepared to bet that—to put it *very* approximately—behavior will prove to be systematic under some of the descriptions under which it is intentional.

At a minimum, the present claim goes like this: there is a way of taxonomizing behaviors and a way of taxonomizing mental states such that, given these taxonomies, theories of the mental causation of behavior will be forthcoming. And that way of taxonomizing mental states construes them nontransparently.

their propositional attitudes. What one would like to say, in particular, is that if two people are identically related to formally identical mental representations, then they are in opaquely type identical mental states. This would be convenient because it yields a succinct and gratifying characterization of what a computational cognitive psychology is about: such a psychology studies propositional attitudes opaquely taxonomized.

I think, in fact, that this is *roughly* the right thing to say, since what I think is *exactly* right is that the construal of propositional attitudes which such a psychology renders is nontransparent. (It's nontransparency that's crucial in all the examples we have been considering). The trouble is that nontransparency isn't quite the same notion as opacity, as we shall now see.

The question before us is: 'What are the relations between the pretheoretic notion of type identity of mental states opaquely construed and the notion of type identity of mental states that you get from a theory which strictly honors the formality condition?' And the answer is: complicated. For one thing, it's not clear that we have *a* pretheoretic notion of the opaque reading of a propositional attitude ascription: I doubt that the two standard tests for opacity (failure of existential generalization and failure of substitutivity of identicals) so much as pick out the same class of cases. But what's more important are the following considerations. While it's notorious that extensionally identical thoughts may be opaquely type distinct (e.g. thoughts about the Morning Star and thoughts about the Evening star) there are nevertheless some semantic conditions on opaque type identification. In particular:

(a) there are some cases of formally distinct but coextensive token thoughts which count as tokens of the same (opaque) type (and hence as identical in content at least on one way of individuating contents); and

(b) *non*-coextensive thoughts are *ipso facto* type distinct (and differ in content at least on one way of individuating contents.)

Cases of type (a): (1) I think I'm sick and you think I'm sick. What's running through my head is 'I'm sick'; what's running through your head is 'he's sick'. But we are both having thoughts of the same (opaque) type (and hence of the same content.)

(2) You think: 'that one looks edible'; I think: 'this one looks edible.' Our thought are opaquely type identical if we are thinking about the same one.

It connects with the existence of such cases that pronouns and demonstratives are typically (perhaps invariably) construed as referring, even when they occur in what are otherwise opaque constructions. So, for example, it seems to me that I can't report Macbeth's hallucination by saying: "Macbeth thinks that's a dagger' if Macbeth is staring at nothing at all. Which is to say that "that's a dagger" doesn't report Macbeth's mental state even though "that's a dagger may be precisely what is running through Macbeth's head (precisely the representation his relation to which is constitutive of his belief).

Cases of type (b): (1) Suppose that Sam feels faint and Misha knows he does. Then what's running through Misha's head may be 'he feels faint.' Suppose too that Misha feels faint and Alfred knows he does. Then what's running through Alfred's head, too, may be 'he feels faint.' I have no, or rather no univocal, inclination to say, in this case, that Alfred and Misha are having type identical thoughts even though the principle of type individuation is, by assumption opaque and even though Alfred and Misha have the same things running through their heads. But if this is right, then formal identity of mental representations cannot be sufficient for type identity of opaquely taxonomized mental states.[8] (There is an interesting discussion of this sort of case in Geach

8. One might try saying: what counts for opaque type individuation is what's *in* your head, not just what's running through it. So, for example, though Alfred and Misha are both thinking, 'he feels faint,' nevertheless different counterfactuals are true of them: Misha would cash his pronoun as: 'he, Sam', whereas Alfred would cash *his* pronoun as: 'he, Misha.' The problem would then be to decide *which* such counterfactuals are relevant, since, if we count all of them, it's going to turn out that there are few, if any, cases of distinct organisms having type identical thoughts.

I won't, in any event, pursue this proposal, since it seems clear that it won't, in principle, cope with all the relevant cases. Two people would be having different thoughts when each is thinking, 'I'm ill' even if *everything* in their heads were the same.

(1957). Geach says that Aquinas says that there is no 'intelligible difference' between Alfred's thought and Misha's. I don't know whether this means that they are having the same thought or that they aren't.)

(2) Suppose that there are two Lake Eries (two bodies of water so-called). Consider two tokens of the thought 'Lake Erie is wet,' one of which is, intuitively speaking, about the Lake Erie in North America and one of which is about the other one. Here again, I'm inclined to say that the aboriginal, uncorrupted, pretheoretical notion of type-wise same thought wants these to be tokens of *different* thoughts and takes these thoughts to differ in content. In this case, though, as in the others, I think there's also a countervailing inclination to say that they count as type identical—and as identical in content—for some relevant purposes and in some relevant respects. How like aboriginal, uncorrupted, pretheoretical intuition!

I think, in short, that the intuitive opaque taxonomy is actually what you might call 'semi-transparent'. On the one hand, certain conditions on coreference are in force (Misha's belief that he's ill is type distinct from Sam's belief that *he's* ill, and my thought *this is edible* may be type identical to your thought *that is edible*. On the other hand, you don't get free substitution of coreferring expressions (beliefs about the Morning Star are type distinct from beliefs about the Evening Star) and existential generalization doesn't go through for beliefs about Santa Claus.

Apparently, then, the notion of same mental state that we get from a theory which honors the formality condition is related to, but not identical to, the notion of same mental state that unreconstructed intuition provides for opaque construals. And it would certainly be reasonable to ask whether we actually need both. I think the answer is probably: yes, if we want to capture *all* the intuitions. For, if we restrict ourselves to either one of the taxonomies we get consequences that we don't like. On the one hand, if we taxonomize *purely* formally, we get identity of belief compatible with difference of truth value. (Misha's belief that he's ill will be type identical to Sam's belief that *he's* ill, but one may be true while the other is false.) On the other hand, if we taxonomize solely according to the pretheoretic criteria, we get trouble with the idea that people act out of their beliefs and desires. We

need, in particular, some taxonomy according to which Sam and Misha have the *same* belief in order to explain why it is that they exhibit the same behaviors. It is, after all, *part* of the pretheoretic notion of belief that difference in belief ought *ceteris paribus* to show up in behavior *somewhere* (*'ceteris paribus'* here means 'given relevant identities among other mental states'), whereas, it's possible to construct cases where differences like the one between Misha's belief and Sam's can't show up in behavior even in principle (see note 8, above). What we have, in short, is a tension between a partially semantic taxonomy and an entirely functional one, and the recommended solution is to use both.

Having said all this, I now propose largely to ignore it and use the term 'opaque taxonomy' for principles of type individuation according to which Misha and Sam are in the same mental state when each believes himself to be ill. When I need to distinguish this sense of opaque taxonomy from the pretheoretic one, I'll talk about *full* opacity and fully opaque type identification.

My claim has been that, in doing our psychology, we want to attribute mental states fully opaquely because it's the fully opaque reading which tells us what the agent has in mind, and it's what the agent has in mind that causes his behavior. I now need to say something about how, precisely, all this is supposed to constitute an argument for the formality condition.

Point one: it's just as well that it's the fully opaque construal of mental states that we need since, patently, that's the only one that the formality condition permits us. This is because the formality condition prohibits taxonomizing psychological states by reference to the semantic properties of mental representations and, at bottom, transparency is a semantic (viz. nonformal; viz. nonsyntactic) notion. The point is sufficiently obvious: if we count the belief that the Evening Star is F as (type) identical to the belief that the Morning Star is F, that must be because of the coreference of such expressions as 'The Morning Star' and 'The Evening Star'. But coreference is a semantic property, and not one which could conceivably have a formal Doppelgänger; it's inconceivable, in particular, that there should be a system of mental representations such that, in the general case, coreferring expressions are formally identical in that system. (This might be true for God's mind, but not, surely, for anybody else's (and

not for God's either unless he is an Extensionalist; which I doubt.)) So, if we want transparent taxonomies of mental states, we will have to give up the formality condition. So it's a good thing for the computational theory of the mind that it's not transparent taxonomies that we want.

What's harder to argue for (but might, nevertheless, be true) is point two: that the formality condition *can* be honored by a theory which taxonomizes mental states according to their content. For, barring caveats previously reviewed, it may be that mental states are distinct in content only if they are relations to formally distinct mental representations; in effect, that aspects of content can be reconstructed as aspects of form, at least insofar as appeals to content figure in accounts of the mental causation of behavior. The main thing to be said in favor of this speculation is that it allows us to explain, within the context of the representational theory of mind, how beliefs of different content *can* have different behavioral effects, even when the beliefs are transparently type identical. The form of explanation goes: it's because different content implies formally distinct internal representations (via the formality condition) and formally distinct internal representations can be functionally different—can differ in their causal role. Whereas, to put it mildly, it's hard to see how internal representations could differ in causal role *unless* they differed in form.

To summarize: transparent taxonomy is patently incompatible with the formality condition; whereas taxonomy in respect of content *may* be compatible with the formality condition, plus or minus a bit. That taxonomy in respect of content *is* compatible with the formality condition, plus or minus a bit, is perhaps *the* basic idea of modern cognitive theory. The representational theory of mind and the computational theory of mind merge here for, on the one hand, it's claimed that psychological states differ in content only if they are relations to type-distinct mental representations; and, on the other, it's claimed that only formal properties of mental representations contribute to their type individuation for the purposes of theories of mind/body interaction. Or, to put it the other way 'round, it's allowed that mental representations affect behavior in virtue of their content, but it's maintained that mental representations are distinct in content only if they are also distinct in form. The first clause is required to make it plausible

that mental states are relations to mental representations and the second is required to make it plausible that mental processes are computations. (Computations just *are* processes in which representations have their causal consequences in virtue of their form.) By thus exploiting the notions of content and computation *together*, a cognitive theory seeks to connect the *intensional* properties of mental states with their *causal* properties vis-à-vis behavior. Which is, of course, exactly what a theory of the mind ought to do.

As must be evident from the preceding, I'm partial to programmatic arguments: ones which seek to infer the probity of a conceptual apparatus from the fact that it plays a role in some prima facie plausible research enterprise. So, in particular, I've argued that a taxonomy of mental states which honors the formality condition seems to be required by theories of the mental causation of behavior, and that that's a reason for taking such taxonomies very seriously.

But there lurks, within the general tradition of representational theories of mind, a deeper intuition: that it is not only *advisable* but actually *mandatory* to assume that mental processes have access only to formal (non-semantic) properties of mental representations; that the contrary view is not only empirically fruitless but also conceptually unsound. I find myself in sympathy with this intuition, though I'm uncertain precisely how the arguments ought to go. What follows is just a sketch.

I'll begin with a version that I *don't* like; an epistemological version.

> Look, it makes no *sense* to suppose that mental operations could apply to mental representations in virtue of (e.g.) the truth or falsity of the latter. For, consider: truth value is a matter of correspondence to the way the world is. To determine the truth value of a belief would therefore involve what I'll call 'directly comparing' the belief with the world; i.e., comparing it with the way the world *is*, not just with the way the world is represented as being. And the representational theory of mind says that we have access to the world only *via* the ways in which we represent it. There is, as it were, nothing that corresponds to looking around (behind? through? what's the right metaphor?) one's beliefs to catch a glimpse of the things they represent.

Mental processes can, in short, compare representations, but they can't compare representations with what they're representations of. Hence mental processes can't have access to the truth value of representations or, *mutatis mutandis*, to whether they denote. Hence the formality condition.

This line of argument could certainly be made a good deal more precise. It has been in, for example, some of the recent work of Nelson Goodman (see especially Goodman, 1978). For present purposes, however, I'm content to leave it *im*precise so long as it sounds familiar. For, I suspect that all versions of the argument suffer from a common deficiency: they assume that you can't run a *correspondence* theory of truth together with a *coherence* theory of evidence. Whereas, I see nothing compelling in the inference from 'truth is a matter of the correspondence of a belief with the way the world is' to '*ascertaining* truth is a matter of "directly comparing" a belief with the way the world is.' Perhaps we ascertain the truth of our beliefs by comparing them with one another, appealing to inference to the best explanation whenever we need to do so.

Anyhow, it would be nice to have a *non*-epistemological defence of the formality condition; one which saves the intuition that there's something conceptually wrong with its denial but doesn't acquire the skeptical/relativistic commitments with which the traditional epistemic versions of the argument have been encumbered. Here goes:

Suppose, just for convenience, that mental processes are algorithms. So, we have rules for the transformation of mental representations, and we have the mental representations which constitute their ranges and domains. Think of the rules as being like hypothetical imperatives; they have antecedents which specify conditions on mental representation, and they have consequents which specify what is to happen if the antecedents are satisfied. And now consider rules *a* and *b*.

(a) If it's the case that P, do such and such.

(b) If you believe it's the case that P, do such and such.

Notice, to begin with, that the compliance conditions on these injunctions are quite different. In particular, in the case where P is *false but believed true*, compliance with *b* consists in doing

such and such, whereas compliance with *a* consists in *not* doing it. But despite this difference in compliance conditions, there's something *very* peculiar (perhaps *pragmatically* peculiar, whatever precisely that may mean) about supposing that an organism might have different ways of going about attempting to comply with *a* and *b*. The peculiarity is patent in *c*:

(c) Do such and such if it's the case that P, *whether or not* you believe that it's the case that P.[9]

To borrow a joke from Professor Robert Jagger, *c* is a little like the advice: 'buy low, sell high.' One knows just what it would be *like* to comply with either, but somehow knowing that doesn't help much.

The idea is this: when one has done what one can to establish that the belief that P is warranted, one has done what one can to establish that the antecedent of *a* is satisfied. And, conversely, when one has done what one can do to establish that the antecedent of *a* is satisfied, one has done what one can to establish the warrant of the belief that P. Now, I suppose that the following is at least *close* to being true: to have the belief that P is to have the belief that the belief that P is warranted; and conversely, to have the belief that the belief that P is warranted is to have the belief that P. And the upshot of *this* is just the formality condition all over again. Given that mental operations have access to the fact that P is believed (and hence that the belief that P is believed to be warranted, and hence that the belief that the belief that P is warranted is believed to be warranted, . . . etc.) there's nothing further left to do; there is nothing that corresponds to the notion of a mental operation which one undertakes to perform just in case one's belief that P is *true*.

This isn't, by the way, any form of skepticism, as can be seen from the following: there's nothing wrong with Jones having one mental operation which he undertakes to perform if it's the case that P and another *quite different* mental operation which he undertakes to perform if *Smith* (≠ Jones) believes that it's the case that P. (Cf. 'I promise . . . though I don't intend to . . .' vs.

9. I'm assuming, for convenience, that all the Ps are such that either they or their denials are believed. This saves having to relativize to time (e.g. having *b* and *c* read '. . . you believe or come to believe . . .').

'I promise . . . though Smith doesn't intend to . . .'). There's a first person/third person asymmetry here, but it doesn't impugn the semantic distinction between 'P is true' and 'P is believed true.' The suggestion is that it's the tacit recognition of this pragmatic asymmetry that accounts for the traditional hunch that you can't both identify mental operations with transformations on mental representations and at the same time flout the formality condition; that the representational theory of mind and the computational theory of mind are somehow conjoint options.

So much, then, for the formality condition and the psychological tradition which accepts it. What about Naturalism? The first point is that none of the arguments *for* a rational psychology is, in and of itself, an argument *against* a Naturalistic psychology. As I remarked above, to deny that mental operations have access to the semantic properties of mental representations is *not* to deny that mental representations *have* semantic properties. On the contrary, beliefs are *just* the kinds of things which exhibit truth and denotation, and the Naturalist proposes to make science out of the organism/environment relations which (presumably) fix these properties. Why, indeed, should he not?

This all *seems* very reasonable. Nevertheless, I now wish to argue that a computational psychology is the only one that we are likely to get; that qua research strategy, the attempt to construct a *naturalistic* psychology is very likely to prove fruitless. I think that the basis for such an argument is already to be found in the literature, where it takes the form of a (possibly inadvertent) reductio ad absurdum of the contrary view.

Consider, to begin with, a distinction that Professor Hilary Putnam introduces in "The Meaning of 'Meaning' " (1975a) between what he calls "psychological states in the wide sense" and "psychological states in the narrow sense". A psychological state in the *narrow* sense is one the ascription of which does not"[presuppose] the existence of any individual other than the subject to whom that state is ascribed" (p. 136). All others are psychological states in the wide sense. So, for example, *x's jealousy of y* is a schema for expressions that denote psychological states in the wide sense, since such expressions presuppose the existence not only of the *x*s who are in the states, but also of the *y*s who are its objects. Putnam remarks that methodological solipsism (the

phrase, by the way, is his) can be viewed as the requirement that only psychological states in the narrow sense are allowed as constructs in psychological theories.

But it is perhaps Putnam's main point that there are at least *some* scientific purposes (e.g. semantics and accounts of intertheoretical reference) which demand the wide construal. Here, rephrased slightly, is the sort of example that Putnam finds persuasive.

There is a planet (call it 'Yon') where things are very much as they are here. In particular, by a cosmic accident, some of the people on Yon speak a dialect indistinguishable from English and live in an urban conglomerate indistinguishable from the Greater Boston Area. Still more, for every one of our Greater Bostonians, there is a Doppelgänger on Yon who has precisely the same neurological structure down to and including microparticles. We can assume that so long as we're construing 'psychological state' narrowly, this latter condition guarantees type identity of our psychological states with theirs.

However, Putnam argues, it doesn't guarantee that there is a corresponding identity of psychological states, hither and Yon, if we construe 'psychological state' *widely*. Suppose that there is this difference between Yon and Earth; whereas, over here, the stuff we call 'water' has the atomic structure H_2O, it turns out that the stuff that they call 'water' over there has the atomic structure XYZ ($\neq H_2O$). And now, consider the mental state *thinking about water*. The idea is that, so long as we construe that state widely, it's one that we, but not our Doppelgängers, can reasonably aspire to. For, construed widely, one is thinking about water only if it is water that one is thinking about. But it's water that one's thinking about only if it is H_2O that one's thinking about; water *is* H_2O. But since, by assumption, they never think about H_2O over Yon, if follows that there's at least one wide psychological state that we're often in and they never are, however neurophysiologically like us they are, and however much our narrow psychological states converge with theirs.

Moreover, if we try to say what they speak about, refer to, mention, etc.—if, in short, we try to supply a semantics for their dialect—we will have to mention XYZ, not H_2O. Hence it would be wrong, at least on Putnam's intuitions, to say that they have a

word for water. A fortiori, the chemists who work in what they call 'M.I.T.' don't have theories about *water*, even though what runs through their head when they talk about XYZ may be identical to what runs through our heads when we talk about H_2O. The situation is analogous to the one which arises for demonstratives and token reflexives, as Putnam insightfully points out.

Well, what are we to make of this? Is it an argument against methodological solipsism? And, if so, is it a *good* argument against methodological solipsism?

To begin with, Putnam's distinction between psychological states in the narrow and wide sense looks to be very intimately related to the traditional distinction between psychological state ascriptions opaquely and transparently construed. I'm a bit wary about this, since what Putnam *says* about wide ascriptions is only that they "presuppose the existence" of objects other than the ascribee; and, of course *a believes Fb and b exists* does not entail *b is such that a believes F of him*, or even $\exists x$ *(a believes Fx)*. Moreover, the failure of such entailments is notoriously important in discussions of quantifying in. For all that, however, I don't *think* that it's Putnam's intention to exploit the difference between the existential generalization test for transparency and the presupposition of existence test for wideness. On the contrary, the burden of Putnam's argument seems to be precisely that 'John believes (widely) that water is F' is true only if water (viz. H_2O) is such that John believes it's F. It's thus unclear to me why Putnam gives the weaker condition on wideness when it appears to be the stronger one that does the work.[10]

But whatever the case may be with the wide sense of belief, it's pretty clear that the narrow sense must be (what I've been calling) fully opaque. This is because it is only full opacity which allows type identity of beliefs that have different truth conditions (Sam's belief that he's ill with Misha's belief that *he* is; Yon beliefs about XYZ with hither beliefs about H_2O). I want to emphasize this correspondence between narrowness and full opacity, and not just in aid of terminological parsimony. Putnam sometimes writes as though he takes the methodological commitment

10. I blush to admit that I had missed some of these complexities until Sylvain Bromberger kindly rubbed my nose in them.

to a psychology of narrow mental states to be a sort of vulgar prejudice: "Making this assumption is, of course, adopting a *restrictive program*—a program which deliberately limits the scope and nature of psychology to fit certain mentalistic preconceptions or, in some cases, to fit an idealistic reconstruction of knowledge and the world" (p. 137). But in light of what we've said so far, it should be clear that this is a methodology with malice aforethought. Narrow psychological states are those individuated in light of the formality condition; viz. without reference to such semantic properties as truth and reference. And honoring the formality condition is part and parcel of the attempt to provide a theory which explains (a) how the belief that the Morning Star is F could be different from the belief that the Evening Star is F despite the well-known astronomical facts; and (b) how the behavioral effects of believing that the Morning Star is F could be different from those of believing that the Evening Star is F, astronomy once again apparently to the contrary notwithstanding. Putnam is, of course, dubious about this whole project: "The three centuries of failure of mentalistic psychology is tremendous evidence against this procedure, in my opinion" (p. 137). I suppose this is intended to include everybody from Locke and Kant to Freud and Chomsky. I should have such failures.

So much for background. I now need an argument to show that a naturalistic psychology (a psychology of mental states transparently individuated; hence, presumably, a psychology of mental states in the wide sense) is, for practical purposes, out of the question. So far as I can see, however, Putnam has given that argument. For, consider: a naturalistic psychology is a theory of organism/environment transactions. So, to stick to Putnam's example, a naturalistic psychology would have to find some stuff S and some relation R, such that one's narrow thought that water is wet is a thought about S in virtue of the fact that one bears R to S. Well, *which* stuff? The natural thing to say would be: 'Water, of course.' Notice, however, that if Putnam is right, it may not even be *true* that the narrow thought that water is wet is a thought about water; it *won't* be true of tokens of that thought which occur on Yon. Whether the narrow thought that water is wet is about water depends on whether it's about H_2O; and whether it's about H_2O depends on 'how science turns out'—viz. on what *chemistry* is

true. (Similarly, mutatis mutandis, *'water' refers to water* is *not*, on this view, a truth of any branch of linguistics; it's *chemists* who tell us what it is that 'water' refers to.) Surely, however, characterizing the objects of thought is methodologically prior to characterizing the causal chains that link thoughts to their objects. But the theory which characterizes the objects of thought is the theory of *everything*; it's all of science. Hence, the methodological moral of Putnam's analysis seems to be: the naturalistic psychologists will inherit the Earth, but only after everybody else is finished with it. No doubt it's alright to have a research strategy that says 'wait awhile'. But who wants to wait *forever?*

This sort of argument isn't novel. Indeed, it was anticipated by Bloomfield (1933). Bloomfield argues that, for all practical purposes, you can't do semantics. The reason you can't is that to do semantics you have to be able to say, for example, what 'salt' refers to. But what 'salt' refers to is NaCl, and that's a bit of chemistry, not linguistics:

> The situations which prompt people to utter speech include every object and happening in their universe. In order to give a scientifically accurate definition of meaning for every form of a language, we would have to have a scientifically accurate knowledge of everything in the speaker's world. The actual extent of human knowledge is very small compared to this. We can define the meaning of a speech-form accurately when this meaning has to do with some matter of which we possess scientific knowledge. We can define the names of minerals, as when we say that the ordinary meaning of the English word *salt* is 'sodium chloride (NaCl),' and we can define the names of plants or animals by means of the technical terms of botany or zoology, but we have no precise way of defining words like *love* or *hate*, which concern situations that have not been accurately classified . . . The statement of meanings is therefore the weak point in language-study, and will remain so until knowledge advances very far beyond its present state.
>
> (pp. 139–140)

It seems to me as though Putnam ought to endorse all of this *including the moral:* the distinction between wanting a

naturalistic semantics (psychology) and not wanting any is real but academic.[11]

The argument just given depends, however, on accepting Putnam's analysis of his example. But suppose that one's intuitions run the other way. Then one is at liberty to argue like this:

1. They do too have water over Yon; all Putnam's example shows is that there could be two kinds of water, our kind (=H_2O) and their kind (=XYZ).
2. Hence, Yon tokens of the thought that water is wet are thoughts about water after all;
3. Hence, the way chemistry turns out is irrelevant to whether thoughts about water are about water.
4. Hence, the naturalistic psychology of thought need not wait upon the sciences of the objects of thought;
5. Hence, a naturalistic psychology may be in the cards after all.

Since the premises of this sort of reply may be tempting (since, indeed, they may be *true*) it's worth presenting a version of the argument which doesn't depend on intuitions about what XYZ is.

A naturalistic psychology would specify the relations that hold between an organism and an object in its environment when the one is thinking about the other. Now, think how such a theory would have to go. Since it would have to define its generalizations over mental states on the one hand and environmental entities on the other, it will need, in particular, some canonical way of referring to the latter. Well, *which* way? If one assumes that what makes my thought about Robin Roberts a thought *about Robin Roberts* is some causal connection between the two of us, then

11. It may be that Putnam *does* accept this moral. For example, the upshot of the discussion circa p. 153 of his article appears to be that a Greek semanticist prior to Archimedes *could* not (in practice) have given a correct account of what (the Greek equivalent of) 'gold' means—because the theory needed to specify the extension of the term was simply not available. Presumably *we* are in that situation vis-à-vis the objects of many of *our* thoughts and the meanings of many of our terms; and, presumably, we will continue to be so into the indefinite future. But then, what's the point of so defining psychology (semantics) that there can't be any?

we'll need a description of RR such that the causal connection obtains in virtue of his satisfying that description. And *that* means, presumably, that we'll need a description under which the relation between him and me instantiates a law.

Generally, then, a naturalistic psychology would attempt to specify environmental objects in a vocabulary such that environment/organism relations are law-instantiating when so described. But here's the depressing consequence again: we have no access to such a vocabulary prior to the elaboration (completion?) of the nonpsychological sciences. 'What Granny likes with her herring' isn't, for example, a description under which salt is law-instantiating; nor, presumably, is 'salt'. What we need is something like 'NaCl', and descriptions like 'NaCl' are available only *after* we've done our chemistry. What this comes down to is that, at a minimum, '*x*'s being F causally explains . . .' can be true only when 'F' expresses nomologically necessary properties of the *x*s. Heaven knows it's hard to say what *that* means, but it presumably rules out both 'Salt's being what Granny likes with herring . . .' and 'Salt's being salt . . . ;' the former for want of being necessary, and the latter for want of being nomological. I take it, moreover, that Bloomfield is right when he says (a) that we don't know relevant nomologically necessary properties of most of the things we can refer to (think about) and (b) that it isn't the linguist's (psychologist's) job to find them out.

Here's still another way to put this sort of argument. The way Bloomfield states his case invites the question: "Why *should* a semanticist want a definition of 'salt' that is "scientifically accurate" in your sense? Why wouldn't a 'nominal' definition do?' There is, I think, some point to such a query. For example, as Hartry Field has pointed out (1972), it wouldn't make much difference to the way that truth-conditional semantics goes if we were to say only " 'salt' refers to whatever it refers to". All we need for this sort of semantics is some way or other of referring to the extension of 'salt'; we don't, in particular, need a "scientifically accurate" way. It's therefore pertinent to do what Bloomfield notably does not: distinguish between the goals of *semantics* and those of a naturalistic psychology of language. The latter, by assumption, purports to explicate the organism/environment transactions in virtue of which relations like reference hold. It

therefore requires, at a minimum, lawlike generalizations of the (approximate) form: *X's utterance of 'salt' refers to salt if X bears relation R to* Δ. Since this whole thing *is* supposed to be lawlike, what goes in for 'Δ' must be a projectible characterization of the extension of 'salt'. But in general we discover which descriptions are projectible only a posteriori, in light of how the sciences (including the nonpsychological sciences) turn out. We are back where we started. Looked at this way, the moral is that we can do (certain kinds of) semantics if we have a way of referring to the extension of 'salt'. But we can't do the naturalistic psychology of reference unless we have some way of saying what salt *is;* which of its properties determine its causal relations.

It's important to emphasize that these sorts of arguments do *not* apply against the research program embodied in 'Rational psychology'—viz. to the program that envisions a psychology that honors the formality condition. The problem we've been facing is: under what description does the object of thought enter into scientific generalizations about the relations between thoughts and their objects? It looks as though the naturalist is going to have to say: under a description that is law instantiating—e.g. under physical description. But the rational psychologist has a quite different answer. What *he* wants is *whatever description the organism has in mind* when it thinks about the object of thought, construing 'thinks about' fully opaquely. So for a theory of psychological states narrowly construed, we want such descriptions of Venus as, e.g., 'The Morning Star', 'The Evening Star', 'Venus', etc., for it is these sorts of descriptions which we presumably entertain when we think that the Morning Star is *F*. In particular, it is our relation to these sorts of descriptions which determine what psychological state type we're in insofar as the goal in taxonomizing psychological states is explaining how they affect behavior.

Final point under the general head: the hopelessness of naturalistic psychology. Practicing naturalistic psychologists have been at least dimly aware all along of the sort of bind that they're in. So, for example, the 'physical specification of the stimulus' is just about invariably announced as a requirement upon adequate formulations of S-R generalizations. We can now see why. Suppose, wildly contrary to fact, that there exists a human population

(e.g. English speakers) in which pencils are, in the technical sense of the notion, discriminative stimuli controlling the verbal response 'pencil'. The point is that even if some such generalization were true, it wouldn't be among those enunciated by a naturalistic psychology; the generalizations of naturalistic psychology are presumably supposed to be nomological, and there aren't any *laws* about pencils *qua* pencils. That is, expressions like 'pencil' presumably occur in no true, lawlike sentences. Of course, there presumably is *some* description in virtue of which pencils fall under the organism/environment laws of a naturalistic psychology, and everybody (except, possibly, Gibson) has always assumed that those descriptions are, approximately, physical descriptions. Hence, the naturalist's demand, perfectly warranted by his lights, that the stimulus should be physically specified.

But though their theory has been consistent, their practice has uniformly not. In practice, and barring the elaborately circumscribed cases that psychophysics studies, the requirement that the stimulus be physically specified has been ignored by just about *all* practitioners. And, indeed, they were well advised to ignore it; how else could they get on with their job? If they really had to wait for the physicists to determine the descriptions(s) under which pencils are law-instantiators, how would the psychology of pencils get off the ground?

So far as I can see, there are really only two ways out of this dilemma:

1. We can fudge, the way that learning theorists usually do. That is, we can 'read' the description of the stimulus from the character of the organism's response. In point of historical fact, this has led to a kind of naturalistic psychology which is merely a solemn paraphrase of what everybody's grandmother knows: e.g. to saying 'pencils are discriminative stimuli for the utterance of "pencil" ' where Granny would have said 'pencil' refers to pencils. I take it that Chomsky's review of *Verbal Behavior* demonstrated, once and for all, the fatuity of this course. What *would* be interesting—what would have surprised Grandmother—is a generalization of the form Δ *is the discriminative stimulus for utterances of 'pencil'* where Δ is a description that picks out pencils in some projectable vocabulary (e.g. in the vocabulary of physics). Does anybody suppose that such descriptions are likely to be forthcoming in, say, the *next* three hundred years?

2. The other choice is to try for a computational psychology—which is, of course, the burden of my plaint. On this view, what we can reasonably hope for is a theory of mental states fully opaquely type individuated. We can try to say what the mental representation is, and what the relation to a mental representation is, such that one believes that the Morning Star is F in virtue of bearing the latter to the former. And we can try to say how that representation, or that relation, or both, differ from the representation and the relation constitutive of believing that the Evening Star is F. A naturalistic psychology, by contrast, remains a sort of ideal of pure reason; there must *be* such a psychology, since, presumably, we do sometimes think of Venus and, presumably, we do so in virtue of a causal relation between it and us. But there's no practical hope of making science out of this relation. And, of course, for methodology, practical hope is *everything.*

One final point, and then I'm through. Methodological solipsism isn't, of course, solipsism *tout court.* It's not part of the enterprise to assert, or even suggest, that you and I are actually in the situation of Winograd's computer. Heaven only knows what relation between me and Robin Roberts makes it possible for me to think of him (refer to him, etc.), and I've been doubting the practical possibility of a science whose generalizations that relation instantiates. But I *don't* doubt that there *is* such a relation or that I do sometimes think of him. Still more: I have reasons not to doubt it; precisely the sorts of reasons I'd supply if I were asked to justify my knowledge claims about his pitching record. In short: it's true that Roberts won 28 and it's true that I know that he did, and nothing in the preceding tends to impugn these truths. (Or, contrariwise, if he didn't and I'm mistaken, then the reasons for my mistake are philosophically boring; they're biographical, not epistemological or ontological.) My point, then, is *of course* not that solipsism is true; it's just that truth, reference, and the rest of the semantic notions aren't psychological categories. What they are is: they're modes of *Dasein.* I don't know what *Dasein* is, but I'm sure that there's lots of it around, and I'm sure that you and I and Cincinnati have all got it. What more do you want?

Acknowledgments: I've had a lot of help with this one. I'm particularly indebted to Professors Ned Block, Sylvain Bromberger, Janet Dean Fodor, Keith Gundersen, Robert Richardson, and Judith Thomson; and to Mr. Israel Krakowski.

12

The Material Mind

DONALD DAVIDSON

I WISH TO DISCUSS some general methodological questions about the nature of psychology as a science by assuming we know very much more than we do about the brain and the nervous system of man. Suppose that we understand what goes on in the brain perfectly, in the sense that we can describe each detail in purely physical terms—that even the electrical and chemical processes, and certainly the neurological ones, have been reduced to physics. And suppose, further, that we see that because of the way the system is constructed, the indeterminacies of quantum physics are irrelevant to our ability to predict and explain the events that are connected with input from sensation or output in the form of motion of the body.

While we are dreaming, let us also dream that the brain, and associated nervous system, have come to be understood as operating much like a computer. We actually come to appreciate what goes on so well that we can build a machine that, when exposed to the lights and sounds of the world, mimics the motions of a man. None of this is absurd, however unlikely or discredited by empirical discoveries it may be.

Finally, partly for fun and partly to stave off questions not germane to the theme, let us imagine that *l'homme machine* has actually been built, in the shape of a man and out of the very stuff of a man, all synthesized from a few dollars' worth of water and other easily obtainable materials. Our evidence that we have