PHILOSOPHY OF MIND

Classical and Contemporary Readings

David J. Chalmers

New York Oxford
OXFORD UNIVERSITY PRESS
2002

Oxford University Press

Oxford New York
Auckland Bangkok Buenos Aires Cape Town Chennai
Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata
Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
São Paulo Shanghai Singapore Taipei Tokyo Toronto

and an associated company in Berlin

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York, 10016
http://www.oup-usa.org

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data
Chalmers, David John, 1966–
Philosophy of mind : classical and contemporary readings / David J. Chalmers.
p. cm.
Includes bibliographical references.
ISBN-13 978-0-19-514581-6
ISBN 0-19-514580-1 (hardback : alk. Paper)—ISBN 0-19-514581-X (pbk : alk.paper)
1. Philosophy of mind. I. Title.

BD418.3 .C435 2002
128′.2—dc21 2002072403

Printing number: 9 8 7 6 5

Printed in the United States of America
on acid-free papers

13 Psychophysical and Theoretical Identifications

David Lewis

Psychophysical identity theorists often say that the identifications they anticipate between mental and neural states are essentially like various uncontroversial theoretical identifications: the identification of water with H_2O, of light with electromagnetic radiation, and so on. Such theoretical identifications are usually described as pieces of voluntary theorizing, as follows. Theoretical advances make it possible to simplify total science by positing bridge laws identifying some of the entities discussed in one theory with entities discussed in another theory. In the name of parsimony, we posit those bridge laws forthwith. Identifications are made, not found.

In 'An Argument for the Identity Theory,'[1] I claimed that this was a bad picture of psychophysical identification, since a suitable physiological theory could *imply* psychophysical identities—not merely make it reasonable to posit them for the sake of parsimony. The implication was as follows:

> Mental state M = the occupant of causal role R (by definition of M).
> Neural state N = the occupant of causal role R (by the physiological theory).
> $\therefore$ Mental state M = neural state N (by transitivity of =).

If the meanings of the names of mental states were really such as to provide the first premise, and if the advance of physiology were such as to provide the second premise, then the conclusion would follow. Physiology and the meanings of words would leave us no choice but to make the psychophysical identification.

In this sequel, I shall uphold the view that psychophysical identifications thus described would be like theoretical identifications, though they would not fit the usual account thereof. For the usual account, I claim, is wrong; theoretical identifications *in general* are implied by the theories that make them possible—not posited independently. This follows from a general hypothesis about the meanings of theoretical terms: that they are definable functionally, by reference to causal roles.[2] Applied to commonsense psychology—folk science rather than professional science, but a theory nonetheless—we get the hypothesis of my previous paper[3] that a mental state M (say, an experience) is definable as the occupant of a certain causal role R—that is, as the state, of whatever sort, that is causally connected in specified ways to sensory stimuli, motor responses, and other mental states.

First, I consider an example of theoretical identification chosen to be remote from past philosophizing; then I give my general account of the meanings of theoretical terms and the nature of theoretical identifications; finally I return to the case of psychophysical identity.

I

We are assembled in the drawing room of the country house; the detective reconstructs the crime. That is, he proposes a *theory* designed to be the best explanation of phenomena we have observed: the death of Mr. Body, the blood on the wallpaper, the silence of the dog in the night, the clock seventeen minutes fast, and so on. He launches into his story:

> *X*, *Y* and *Z* conspired to murder Mr. Body. Seventeen years ago, in the gold fields of Uganda, *X* was Body's partner . . . Last week, *Y* and *Z* conferred in a bar in Reading . . . Tuesday night at 11:17, *Y* went to the attic and set a time bomb . . . Seventeen minutes later, *X* met *Z* in the billiard room and gave him the lead pipe . . . Just when the bomb went off in the attic, *X* fired three shots into the study through the French windows . . .

And so it goes: a long story. Let us pretend that it is a single long conjunctive sentence.

The story contains the three names '*X*', '*Y*' and '*Z*'. The detective uses these new terms without explanation, as though we knew what they meant. But we do not. We never used them before, at least not in the senses they bear in the present context. All we know about their meanings is what we gradually gather from the story itself. Call these *theoretical terms* (*T-terms* for

From *Australasian Journal of Philosophy* 50:249–58, 1972. Reprinted with permission of the author's estate and the publisher.

short) because they are introduced by a theory. Call the rest of the terms in the story *O-terms.* These are all the *other* terms except the T-terms; they are all the *old, original* terms we understood before the theory was proposed. We could call them our 'pre-theoretical' terms. But 'O' does *not* stand for 'observational.' Not all the O-terms are observational terms, whatever those may be. They are just any old terms. If part of the story was mathematical—if it included a calculation of the trajectory that took the second bullet to the chandelier without breaking the vase—then some of the O-terms will be mathematical. If the story says that something happened because of something else, then the O-terms will include the intensional connective 'because,' or the operator 'it is a law that,' or something of the sort.

Nor do the theoretical terms name some sort of peculiar theoretical, unobservable, semi-fictitious entities. The story makes plain that they name *people.* Not theoretical people, different somehow from ordinary, observational people—just people!

On my account, the detective plunged right into his story, using '*X*', '*Y*' and '*Z*' as if they were names with understood denotation. It would have made little difference if he had started, instead, with initial existential quantifiers: 'There exist *X, Y* and *Z* such that . . .' and then told the story. In that case, the terms '*X*', '*Y*' and '*Z*' would have been bound variables rather than T-terms. But the story would have had the same explanatory power. The second version of the story, with the T-terms turned into variables bound by existential quantifiers, is the Ramsey sentence of the first. Bear in mind, as evidence for what is to come, how little difference the initial quantifiers seem to make to the detective's assertion.

Suppose that after we have heard the detective's story, we learn that it is true of a certain three people: Plum, Peacock and Mustard. If we put the name 'Plum' in place of '*X*', 'Peacock' in place of '*Y*', and 'Mustard' in place of '*Z*' throughout, we get a true story about the doings of those three people. We will say that Plum, Peacock and Mustard together *realize* (or are a *realization* of) the detective's theory.

We may also find out that the story is not true of any other triple.[4] Put in any three names that do not name Plum, Peacock and Mustard (in that order) and the story we get is false. We will say that Plum, Peacock and Mustard *uniquely realize* (are the *unique realization* of) the theory.

We might learn both of these facts. (The detective might have known them all along, but held them back to spring his trap; or he, like us, might learn them only after his story had been told.) And if we did, we would surely conclude that *X, Y* and *Z* in the story were Plum, Peacock and Mustard. I maintain that we would be compelled so to conclude, given the senses borne by the terms '*X*', '*Y*' and '*Z*' in virtue of the way the detective introduced them in his theorizing, and given our information about Plum, Peacock and Mustard.

In telling his story, the detective set forth three roles and said that they were occupied by *X, Y* and *Z.* He must have specified the meanings of the three T-terms '*X*', '*Y*' and '*Z*' thereby; for they had meanings afterwards, they had none before, and nothing else was done to give them meanings. They were introduced by an implicit functional definition, being reserved to name the occupants of the three roles. When we find out who are the occupants of the three roles, we find out who are *X, Y* and *Z.* Here is our theoretical identification.

In saying that the roles were occupied by *X, Y* and *Z,* the detective implied that they were occupied. That is, his theory implied its Ramsey sentence. That seems right; if we learnt that no triple realized the story, or even came close, we would have to conclude that the story was false. We would also have to deny that the names '*X*', '*Y*' and '*Z*' named anything; for they were introduced as names for the occupants of roles that turned out to be unoccupied.

I also claim that the detective implied that the roles were uniquely occupied, when he reserved names for their occupants and proceeded as if those names had been given definite referents. Suppose we learnt that two different triples realized the theory: Plum, Peacock, Mustard; and Green, White, Scarlet. (Or the two different triples might overlap; Plum, Peacock, Mustard; and Green, Peacock, Scarlet.) I think we would be most inclined to say that the story was false, and that the names '*X*', '*Y*' and '*Z*' did not name anything. They were introduced as names for the occupants of certain roles; but there is no such thing as *the* occupant of a doubly occupied role, so there is nothing suitable for them to name.

If, as I claim, the T-terms are definable as naming the first, second, and third components of the unique triple that realizes the story, then the T-terms can be treated like definite descriptions. If the story is uniquely realized, they

name what they ought to name; if the story is unrealized or multiply realized, they are like improper descriptions. If too many triples realize the story, '*X*' is like 'the moon of Mars'; if too few triples—none—realize the story, '*X*' is like 'the moon of Venus.' Improper descriptions are not meaningless. Hilary Putnam has objected that on this sort of account of theoretical terms, the theoretical terms of a falsified theory come out meaningless.[5] But they do not, if theoretical terms of unrealized theories are like improper descriptions. 'The moon of Mars' and 'The moon of Venus' do not (in any normal way) name anything here in our actual world; but they are not meaningless, because we know very well what they name in certain alternative possible worlds. Similarly, we know what '*X*' names in any world where the detective's theory is true, whether or not our actual world is such a world.

A complication: what if the theorizing detective has made one little mistake? He should have said that *Y* went to the attic at 11:37, not 11:17. The story as told is unrealized, true of no one. But another story is realized, indeed uniquely realized: the story we get by deleting or correcting the little mistake. We can say that the story as told is *nearly realized,* has a unique *near-realization.* (The notion of a near-realization is hard to analyze, but easy to understand.) In this case the T-terms ought to name the components of the near-realization. More generally: they should name the components of the nearest realization of the theory, provided there is a unique nearest realization and it is near enough. Only if the story comes nowhere near to being realized, or if there are two equally near nearest realizations, should we resort to treating the T-terms like improper descriptions. But let us set aside this complication for the sake of simplicity, though we know well that scientific theories are often nearly realized but rarely realized, and that theoretical reduction is usually blended with revision of the reduced theory.

This completes our example. It may seem atypical; the T-terms are names, not predicates or functors. But that is of no importance. It is a popular exercise to recast a language so that its nonlogical vocabulary consists entirely of predicates; but it is just as easy to recast a language so that its nonlogical vocabulary consists entirely of names (provided that the logical vocabulary includes a copula). These names, of course, may purport to name individuals, sets, attributes, species, states, functions, relations, magnitudes, phenomena or what have you; but they are still names. Assume this done, so that we may replace all T-terms by variables of the same sort.

II

We now proceed to a general account of the functional definability of T-terms and the nature of theoretical identification. Suppose we have a new theory, *T,* introducing the new terms $t_1 \ldots t_n$. These are our T-terms. (Let them be names.) Every other term in our vocabulary, therefore, is an O-term. The theory *T* is presented in a sentence called the *postulate* of *T.* Assume this is a single sentence, perhaps a long conjunction. It says of the entities—states, magnitudes, species, or whatever—named by the T-terms that they occupy certain *causal roles;* that they stand in specified causal (and other) relations to entities named by O-terms, and to one another. We write the postulate thus:[6]

$$\mathbf{T[t]}.$$

Replacing the T-terms uniformly by free variables $x_1 \ldots x_n$, we get a formula in which only O-terms appear:

$$\mathbf{T[x]}.$$

Any *n*-tuple of entities which satisfies this formula is a realization of the theory *T.* Prefixing existential quantifiers, we get the *Ramsey sentence* of *T,* which says that *T* has at least one realization:

$$\exists \mathbf{x}\ \mathbf{T[x]}.$$

We can also write a *modified Ramsey sentence* which says that *T* has a unique realization:[7]

$$\exists_1 \mathbf{x}\ \mathbf{T[x]}.$$

The Ramsey sentence has exactly the same O-content as the postulate of *T;* any sentence free of T-terms follows logically from one if and only if it follows from the other.[8] The modified Ramsey sentence has slightly more O-content. I claim that this surplus O-content does belong to the theory *T*—there are more theorems of *T* than follow logically from the postulate alone. For in presenting the postulate as if the T-terms has been well-defined thereby, the theorist has implictly asserted that *T* is uniquely realized.

We can write the *Carnap sentence* of *T:* the conditional of the Ramsey sentence and the postulate, which says that if *T* is realized, then the

T-terms name the components of some realization of *T:*

$$\exists \mathbf{x}\, T[\mathbf{x}] \supset T[\mathbf{t}].$$

Carnap has suggested this sentence as a meaning postulate for *T;*[9] but if we want T-terms of unrealized or multiply realized theories to have the status of improper descriptions, our meaning postulates should instead be a *modified Carnap sentence,* this conditional with our modified Ramsey sentence as antecedent:

$$\exists_1 \mathbf{x}\, T[\mathbf{x}] \supset T[\mathbf{t}],$$

together with another conditional to cover the remaining cases:[10]

$$\sim \exists_1 \mathbf{x}\, T[\mathbf{x}] \supset \mathbf{t} = *.$$

This pair of meaning postulates is logically equivalent[11] to a sentence which explicitly defines the T-terms by means of O-terms:

$$\mathbf{t} = \iota \mathbf{x}\, T[\mathbf{x}].$$

This is what I have called functional definition. The T-terms have been defined as the occupants of the causal roles specified by the theory *T;* as *the* entities, whatever those may be, that bear certain causal relations to one another and to the referents of the O-terms.

If I am right, T-terms are eliminable—we can always replace them by their definientia. Of course, this is not to say that theories are fictions, or that theories are uninterpreted formal abacuses, or that theoretical entities are unreal. Quite the opposite! Because we understand the O-terms, and we can define the T-terms from them, theories are fully meaningful; we have reason to think a good theory true; and if a theory is true, then whatever exists according to the theory really *does* exist.

I said that there are more theorems of *T* than follow logically from the postulate alone. More precisely: the theorems of *T* are just those sentences which follow from the postulate together with the corresponding functional definition of the T-terms. For that definition, I claim, is given implicitly when the postulate is presented as bestowing meanings on the T-terms introduced in it.

It may happen, after the introduction of the T-terms, that we come to believe of a certain *n*-tuple of entities, specified otherwise than as the entities that realize *T,* that they do realize *T.* That is, we may come to accept a sentence

$$T[\mathbf{r}]$$

where $r_1 \ldots r_n$ are either O-terms or theoretical terms of some other theory, introduced into our language independently of $t_1 \ldots t_n$. This sentence, which we may call a *weak reduction premise* for *T,* is free of T-terms. Our acceptance of it might have nothing to do with our previous acceptance of *T.* We might accept it as part of some new theory; or we might believe it as part of our miscellaneous, unsystematized general knowledge. Yet having accepted it, for whatever reason, we are logically compelled to make theoretical identifications. The reduction premise, together with the functional definition of the T-terms and the postulate of *T,* logically implies the identity:

$$\mathbf{t} = \mathbf{r}.$$

In other words, the postulate and the weak reduction premise definitionally imply the identities $t_i = r_i$.

Or we might somehow come to believe of a certain *n*-tuple of entities that they *uniquely* realize *T;* that is, to accept a sentence

$$\forall \mathbf{x}(T[\mathbf{x}] \equiv \mathbf{x} = \mathbf{r})$$

where $r_1 \ldots r_n$ are as above. We may call this a *strong reduction premise* for *T,* since it definitionally implies the theoretical identifications by itself, without the aid of the postulate of *T.* The strong reduction premise logically implies the identity

$$\mathbf{r} = \iota \mathbf{x}\, T[\mathbf{x}]$$

which, together with the functional definition of the T-terms, implies the identities $t_i = r_i$ by transitivity of identity.

These theoretical identifications are not voluntary posits, made in the name of parsimony; they are deductive inferences. According to their definitions, the T-terms name the occupants of the causal roles specified by the theory *T.* According to the weak reduction premise and *T,* or the strong reduction premise by itself, the occupants of those causal roles turn out to be the referents of $r_1 \ldots r_n$. Therefore, those are the entities named by the T-terms. That is how we inferred that *X, Y* and *Z* were Plum, Peacock and Mustard; and that, I suggest, is how we make theoretical identifications in general.

III

And that is how, someday, we will infer that[12] the mental states $M_1, M_2, \ldots$ are the neural states $N_1, N_2, \ldots$.

Think of common-sense psychology as a term-introducing scientific theory, though one invented long before there was any such institution as professional science. Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses. Perhaps we can think of them as having the form:

> When someone is in so-and-so combination of mental states and receives sensory stimuli of so-and-so kind, he tends with so-and-so probability to be caused thereby to go into so-and-so mental states and produce so-and-so motor responses.

Add also all the platitudes to the effect that one mental state falls under another—'toothache is a kind of pain,' and the like. Perhaps there are platitudes of other forms as well. Include only platitudes which are common knowledge among us—everyone knows them, everyone knows that everyone else knows them, and so on. For the meanings of our words are common knowledge, and I am going to claim that names of mental states derive their meaning from these platitudes.

Form the conjunction of these platitudes; or better, form a cluster of them—a disjunction of all conjunctions of *most* of them. (That way it will not matter if a few are wrong.) This is the postulate of our term-introducing theory. The names of mental states are the T-terms.[13] The O-terms used to introduce them must be sufficient for speaking of stimuli and responses, and for speaking of causal relations among these and states of unspecified nature.

From the postulate, form the definition of the T-terms; it defines the mental states by reference to their causal relations to stimuli, responses, and each other. When we learn what sort of states occupy those causal roles definitive of the mental states, we will learn what states the mental states are—exactly as we found out who *X* was when we found out that Plum was the man who occupied a certain role, and exactly as we found out what light was when we found that electromagnetic radiation was the phenomenon that occupied a certain role.

Imagine our ancestors first speaking only of external things, stimuli, and responses—and perhaps producing what we, but not they, may call *Aüsserungen* of mental states—until some genius invented the theory of mental states, with its newly introduced T-terms, to explain the regularities among stimuli and responses. But that did not happen. Our common-sense psychology was never a newly invented term-introducing scientific theory—not even of prehistoric folk-science. The story that mental terms were introduced as theoretical terms is a myth.

It is, in fact, Sellars' myth of our Rylean ancestors.[14] And though it is a myth, it may be a good myth or a bad one. It is a good myth if our names of mental states do in fact mean just what they would mean if the myth were true.[15] I adopt the working hypothesis that it is a good myth. This hypothesis can be tested, in principle, in whatever way any hypothesis about the conventional meanings of our words can be tested. I have not tested it; but I offer one item of evidence. Many philosophers have found Rylean behaviorism at least plausible; more have found watered down, 'criteriological' behaviorism plausible. There is a strong odor of analyticity about the platitudes of common-sense psychology. The myth explains the odor of analyticity and the plausibility of behaviorism. If the names of mental states are like theoretical terms, they name nothing unless the theory (the cluster of platitudes) is more or less true. Hence it is analytic that *either* pain, etc., do not exist *or* most of our platitudes about them are true. If this *seems* analytic to you, you should accept the myth, and be prepared for psychophysical identifications.

The hypothesis that names of mental states are like functionally defined theoretical terms solves a familiar problem about mental explanations. How can my behavior be explained by an explanans consisting of nothing but particular-fact premises about my present state of mind? Where are the covering laws? The solution is that the requisite covering laws are implied by the particular-fact premises. Ascriptions to me of various particular beliefs and desires, say, cannot be true if there are no such states as belief and desire; cannot be true, that is, unless the causal roles definitive of belief and desire are occupied. But these roles can only be occupied by states causally related in the proper lawful way to behavior.

Formally, suppose we have a mental explanation of behavior as follows.

$$\frac{C_1[\mathbf{t}], C_2[\mathbf{t}], \ldots}{E}$$

Here E describes the behavior to be explained; $C_1[\mathbf{t}]$, $C_2[\mathbf{t}]$, . . . are particular-fact premises describing the agent's state of mind at the time. Various of the mental terms $t_1 \ldots t_n$ appear in these premises, in such a way that the premises would be false if the terms named nothing. Now

let $L_1[\mathbf{t}], L_2[\mathbf{t}], \ldots$ be the platitudinous purported causal laws whereby—according to the myth—the mental terms were introduced. Ignoring clustering for simplicity, we may take the term-introducing postulate to be the conjunction of these. Then our explanation may be rewritten:

$$\frac{\exists_1 \mathbf{x}\left(\begin{array}{l} L_1[\mathbf{x}] \,\&\, L_2[\mathbf{x}] \,\&\, \ldots \,\&\, \\ C_1[\mathbf{x}] \,\&\, C_2[\mathbf{x}] \,\&\, \ldots \end{array}\right)}{E}$$

The new explanans is a definitional consequence of the original one. In the expanded version, however, laws appear explicitly alongside the particular-fact premises. We have, so to speak, an existential generalization of an ordinary covering-law explanation.[16]

The causal definability of mental terms has been thought to contradict the necessary infallibility of introspection.[17] Pain is one state; belief that one is in pain is another. (Confusingly, either of the two may be called 'awareness of pain.') Why cannot I believe that I am in pain without being in pain—that is, without being in whatever state it is that occupies so-and-so causal role? Doubtless I am so built that this normally does not happen; but what makes it impossible?

I do not know whether introspection is (in some or all cases) infallible. But if it is, that is no difficulty for me. Here it is important that, on my version of causal definability, the mental terms stand or fall together. If common-sense psychology fails, all of them are alike denotationless.

Suppose that among the platitudes are some to the effect that introspection is reliable: 'belief that one is in pain never occurs unless pain occurs' or the like. Suppose further that these platitudes enter the term-introducing postulate as conjuncts, not as cluster members; and suppose that they are so important that an n-tuple that fails to satisfy them perfectly is not even a near-realization of common-sense psychology. (I neither endorse nor repudiate these suppositions.) Then the necessary infallibility of introspection is assured. Two states cannot be pain and belief that one is in pain, respectively (in the case of a given individual or species) if the second *ever* occurs without the first. The state that *usually* occupies the role of belief that one is in pain may, of course, occur without the state that *usually* occupies the role of pain; but in that case (under the suppositions above) the former no longer is the state of belief that one is in pain, and the latter no longer is pain. Indeed, the victim no longer is in any mental state whatever, since his states no longer realize (or nearly realize) common-sense psychology. Therefore it is impossible to believe that one is in pain and not be in pain.

NOTES

Previous versions of this paper were presented at a conference on Philosophical Problems of Psychology held at Honolulu in March, 1968; at the annual meeting of the Australasian Association of Philosophy held at Brisbane in August, 1971; and at various university colloquia. This paper is expected to appear also in a volume edited by Chung-ying Cheng.

1. *Journal of Philosophy,* 63 (1966): 17–25.
2. See my 'How to Define Theoretical Terms,' *Journal of Philosophy,* 67 (1970): 427–446.
3. Since advocated also by D. M. Armstrong, in *A Materialist Theory of the Mind* (New York: Humanities Press, 1968). He expresses it thus: 'The concept of a mental state is primarily the concept of a state of the person apt for bringing about a certain sort of behaviour [and secondarily also, in some cases] apt for being brought about by a certain sort of stimulus,' p. 82.
4. The story itself might imply this. If, for instance, the story said '*X* saw *Y* give *Z* the candlestick while the three of them were alone in the billiard room at 9:17,' then the story could not possibly be true of more than one triple.
5. 'What Theories Are Not,' in Nagel, Suppes and Tarski eds., *Logic, Methodology and Philosophy of Science* (Stanford University Press, 1962): 247.
6. Notation: boldface names and variables denote n-tuples; the corresponding subscripted names and variables denote components of n-tuples. For instance, $\mathbf{t}$ is $<t_1 \ldots t_n>$. This notation is easily dispensable, and hence carries no ontic commitment to n-tuples.
7. That is, $\exists \mathbf{y} \forall \mathbf{x}(T[\mathbf{x}] \equiv \mathbf{y} = \mathbf{x})$. Note that $\exists_1 x_1 \ldots \exists_1 x_n T[\mathbf{x}]$ does not imply $\exists_1 \mathbf{x}\, T[\mathbf{x}]$, and does not say that T is uniquely realized.
8. On the assumptions—reasonable for the postulate of a scientific theory—that the T-terms occur purely referentially in the postulate, and in such a way that the postulate is false if any of them are denotationless. We shall make these assumptions henceforth.
9. Most recently in *Philosophical Foundations of Physics* (New York: Basic Books, 1966): 265–274. Carnap, of course, has in mind the case in which the O-terms belong to an observation language.
10. $\mathbf{t} = *$ means that each t_i is denotationless. Let * be some chosen necessarily denotationless name; then * is $<* \ldots *>$ and $\mathbf{t} = *$ is equivalent to the conjunction of all the identities $t_i = *$.
11. Given a theory of descriptions which makes an iden-

tity true whenever both its terms have the status of improper descriptions, false whenever one term has that status and the other does not. This might best be the theory of descriptions in Dana Scott, 'Existence and Description in Formal Logic,' in R. Schoenman, ed., *Bertrand Russell: Philosopher of the Century* (London: Allen & Unwin, 1967).

12. In general, or in the case of a given species, or in the case of a given person. It might turn out that the causal roles definitive of mental states are occupied by different neural (or other) states in different organisms. See my discussion of Hilary Putnam 'Psychological Predicates' in *Journal of Philosophy,* 66 (1969): 23–25.
13. It may be objected that the number of mental states is infinite, or at least enormous; for instance, there are as many states of belief as there are propositions to be believed. But it would be better to say that there is one state of belief, and it is a relational state, relating people to propositions. (Similarly, centigrade temperature is a relational state, relating objects to numbers.) The platitudes involving belief would, of course, contain universally quantified proposition-variables. Likewise for other mental states with intentional objects.
14. Wilfrid Sellars, 'Empiricism and the Philosophy of Mind,' in Feigl and Scriven, eds., *Minnesota Studies in the Philosophy of Science,* I (University of Minnesota Press, 1956): 309–20.
15. Two myths which cannot both be true together can nevertheless both be good together. Part of my myth says that names of color-sensations were T-terms, introduced using names of colors as O-terms. If this is a good myth, we should be able to define 'sensation of red' roughly as 'that state apt for being brought about by the presence of something red (before one's open eyes, in good light, etc.).' A second myth says that names of colors were T-terms introduced using names of color-sensations as O-terms. If this second myth is good, we should be able to define 'red' roughly as 'that property of things apt for bringing about the sensation of red.' The two myths could not both be true, for which came first: names of color-sensations or of colors? But they could both be good. We could have a circle in which colors are correctly defined in terms of sensations and sensations are correctly defined in terms of colors. We could not discover the meanings *both* of names of colors and of names of color-sensations just by looking at the circle of correct definitions, but so what?
16. See 'How to Define Theoretical Terms': 440–441.
17. By Armstrong, in *A Materialist Theory of the Mind,* pp. 100–13. He finds independent grounds for denying the infallibility of introspection.

Troubles with Functionalism

Ned Block

. . . One characterization of functionalism that is probably vague enough to be accepted by most functionalists is: each type of mental state is a state consisting of a disposition to act in certain ways *and to have certain mental states,* given certain sensory inputs and certain mental states. So put, functionalism can be seen as a new incarnation of behaviorism. Behaviorism identifies mental states with dispositions to act in certain ways in certain input situations. But as critics have pointed out (Chisholm, 1957; Putnam, 1963), desire for goal G cannot be identified with, say, the disposition to do A in input circumstances in which A leads to G, since, after all, the agent might not *know* A leads to G and thus might not be disposed to do A. Functionalism replaces behaviorism's "sensory inputs" with "sensory inputs and mental states"; and functionalism replaces behaviorism's "disposition to act" with "disposition to act and have certain mental states." Functionalists want to individuate mental states causally, and since mental states have mental causes and effects as well as sensory causes and behavioral effects, functionalists individuate mental states partly in terms of causal relations to other mental states. One consequence of this difference between functionalism and behaviorism is that there are organisms that according to behaviorism, have mental states but, according to functionalism, do not have mental states.

So, necessary conditions for mentality that are postulated by functionalism are in one respect stronger than those postulated by behav-

Excerpted from C. W. Savage, ed., *Perception and Cognition* (University of Minnesota Press, 1978), pp. 261–325, with permission of the publisher.