1 Functionalism

What Is Functionalism?

Functionalism is one of the major proposals that have been offered as solutions to the mind-body problem. Solutions to the mind-body problem usually try to answer questions such as: What is the ultimate nature of the mental? At the most general level, what makes a mental state mental? Or more specifically, What do thoughts have in common by virtue of which they are thoughts? That is, what makes a thought a thought? What makes a pain a pain? Cartesian dualism said the ultimate nature of the mental was to be found in a special mental substance. Behaviorism identified mental states with behavioral dispositions; physicalism in its most influential version identifies mental states with brain states. Functionalism says that mental states are constituted by their causal relations to one another and to sensory inputs and behavioral outputs. Although it is descended from Aristotle, modern functionalism is one of the major theoretical developments of twentieth-century analytic philosophy, and provides the conceptual underpinnings of much work in cognitive science.

Functionalism has three distinct modern-day sources. First, Putnam and Fodor saw mental states in terms of an empirical computational theory of the mind. Second, Smart's ''topic-neutral'' analyses led Armstrong and Lewis to a functionalist analysis of mental concepts. Third, Wittgenstein's idea of meaning as use led to a version of functionalism as a theory of meaning, further developed by Sellars and later Harman.

One motivation behind functionalism can be appreciated by attention to artifact concepts like *carburetor* and biological concepts like *kidney*. What it is for something to be a carburetor is for it to mix fuel and air in an internal combustion engine—*carburetor* is a functional concept. In the case of the kidney, the *scientific* concept is functional—defined in terms of a role in filtering the blood and maintaining certain chemical balances.

The kind of function relevant to the mind can be introduced via the parity-detecting automaton illustrated in figure 1.1, which tells us whether it has seen an odd or even number of '1's. This automaton has two states, S_1 and S_2; two inputs, '1' and '0'; and

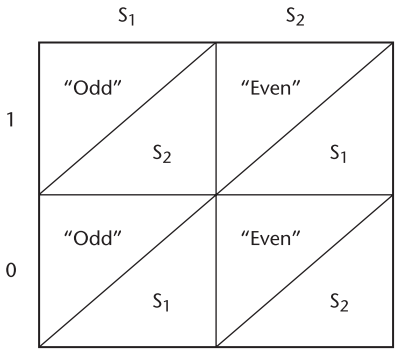

Figure 1.1
Parity automaton with two inputs

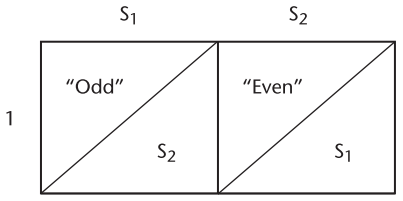

Figure 1.2
Parity automaton with one input

two outputs, with either the word ''Odd'' or ''Even'' being uttered. Figure 1.1 describes two functions, one from input and state to output, and another from input and state to next state. Each square encodes two conditionals specifying the output and next state given both the current state and input. The top-left box says that if the machine is in S_1 and sees a '1', it says ''Odd'' (indicating that it has seen an odd number of '1's) and goes to S_2. The right box says, similarly, that if the machine is in S_2 and sees a '1', it says ''Even'' and goes back to S_1. The bottom-left box says that if the machine is in S_1 and sees a '0', it says ''Odd'' and stays in S_1. The machine is intended to start in S_1, so if its first input is a '0', it will wrongly say that it has seen an odd number of '1's, but once it has seen a 1, subsequent answers will be correct. (The flaw is corrected in the next machine.)

The machine of figure 1.2 is simpler. As before, this automaton has two states, S_1 and S_2, and two outputs, ''Odd'' or ''Even.'' The difference is that it only has one input, '1', though of course it can get no input at all (as can the machine of figure 1.1). As before, the table describes two functions, one from input and state to output, and another from input and state to next state. As before, each square encodes two conditionals

specifying the output and next state given both the current state and input. The left box says that if the machine is in S_1 and sees a '1,' it says "Odd" (indicating that it has seen an odd number of '1's) and goes to S_2. The right box says, similarly, that if the machine is in S_2 and sees a '1', it says "Even" and goes back to S_1. This machine is simpler than the machine of figure 1.1 and intuitively serves the same purpose and further avoids branding zero '1's as an odd number of '1's.

Now suppose we ask the question: "What is S_1?" The answer is that the nature of S_1 is entirely relational, and entirely captured by the figure. We could give an explicit characterization of S_1 (from figure 1.2) as follows:

Being in S_1 = being in the first of two states that are related to one another and to inputs and outputs as follows: being in one of the states and getting a '1' input results in going into the second state and emitting "Odd"; and being in the second of the two states and getting a '1' input results in going into the first and emitting "Even."

Making the quantification over states more explicit:

Being in S_1 = Being an x such that $\exists P \exists Q$ [If x is in P and gets a '1' input, then it goes into Q and emits "Odd"; if x is in Q and gets a '1' input, it goes into P and emits "Even" & x is in P]. (Note: Read "$\exists P$" as There is a property P.)

This illustration can be used to make a number of points: (1) According to functionalism, the nature of a mental state is just like the nature of an automaton state—that is, constituted by its relations to other states and to inputs and outputs. All there is to S_1 is that being in it and getting a '1' input results in such and such, etc. According to functionalism, all there is to being in pain is that it disposes you to say "ouch," wonder whether you are ill, become distracted, and so on. (2) Because mental states are like automaton states in this regard, the illustrated method for defining automaton states is supposed to work for mental states as well. Mental states can be totally characterized in terms that involve only logicomathematical language and terms for input signals and behavioral outputs. Thus functionalism satisfies one of the desiderata of behaviorism, characterizing the mental in entirely nonmental language. (3) S_1 is a second-order state in that it consists in having *other* properties, say mechanical or hydraulic or electronic properties, that have certain relations to one another. These other properties, the ones quantified over in the definitions just given, are said to be the *realizations* of the functional properties. So, although functionalism characterizes the mental in nonmental terms, it does so only by quantifying over realizations of mental states, which would not have delighted behaviorists. (4) One functional state can be realized in different ways. For example, an actual metal and plastic machine satisfying the machine table might be made of gears, wheels, pulleys, and the like, in which case the realization of S_1 would be a mechanical state; or the realization of S_1 might be an electronic state, and so forth. (5) Just as one functional state can be realized in different ways, one

physical state can realize different functional states in different machines. This could happen, for example, if a single type of transistor were used to do different things in different machines. (6) Since S_1 can be realized in many ways, a claim that S_1 *is* a mechanical state would be false (at least arguably), as would a claim that S_1 is an electronic state. For this reason, functionalism shows that physicalism is false: if a creature without a brain can think, thinking cannot be a brain state. (But see the section on functionalism and physicalism below.)

The notion of a realization deserves further discussion. In the early days of functionalism, a first-order property was often said to realize a functional property by virtue of a 1-1 correspondence between the two realms of properties. But such a definition of realization produces far too many realizations. Suppose, for example, that at t_1 we shout "one" at a bucket of water, and then at t_2 we shout "one" again. We can regard the bucket as a parity-detecting automaton by pairing the physical configuration of the bucket at t_1 with S_1 and the heat emitted or absorbed by the bucket at t_1 with "Odd"; by pairing the physical configuration of the bucket at t_2 with S_2 and the heat exchanged with the environment at t_2 with "Even"; and so on. What is left out by the post hoc correlation way of thinking of realization is that a true realization must satisfy the *counterfactuals* implicit in figure 1.2. To be a realization of S_1, it is not enough to lead to a certain output and state given that the input is a '1'; it is also required that had the input been a '0', the S_1 realization would have led to the other output and *state*. Satisfaction of the relevant counterfactuals is built into the notion of realization mentioned in (3) above. See Lycan 1987.

Suppose we have a theory of mental states that specifies all the causal relations among the states, sensory inputs, and behavioral outputs. Focusing on pain as a sample mental state, it might say, among other things, that sitting on a tack causes pain and that pain causes anxiety and saying "ouch." Agreeing to go along with this moronic theory for the sake of the example, functionalism would then say that we could define "pain" as follows: being in pain = being in the first of two states, the first of which is caused by sitting on tacks, and which in turn causes the other state and emitting "ouch." More symbolically

Being in pain = Being an x such that $\exists P \exists Q$ [sitting on a tack causes P & P causes both Q and emitting "ouch" & x is in P]

More generally, if T is a psychological theory with n mental terms of which the seventeenth is "pain," we can define "pain" relative to T as follows. (The 'F_1' ... 'F_n' are variables that replace the n mental terms, and 'i_1', etc., and 'o_1', etc. indicate input and output terms).

Being in pain = Being an x such that $\exists F_1 \ldots \exists F_n$ [$T(F_1 \ldots F_n, i_1$, etc., o_1, etc.) & x is in F_{17}]

In this way, functionalism characterizes the mental in nonmental terms, in terms that involve quantification over realizations of mental states but no explicit mention of them; thus functionalism characterizes the mental in terms of structures that are tacked down to reality only at the inputs and outputs.

The psychological theory *T* just mentioned can be either an empirical psychological theory or else a commonsense ''folk'' theory, and the resulting functionalisms are very different. In the latter case, *conceptual functionalism*, the functional definitions are aimed at capturing our ordinary mental concepts. In the former case, which I named *psychofunctionalism*, the functional definitions are not supposed to capture ordinary concepts but are only supposed to fix the extensions of mental terms. The idea of psychofunctionalism is that the scientific nature of the mental consists not in anything biological, but in something ''organizational,'' analogous to computational structure. Conceptual functionalism, by contrast, can be thought of as a development of logical behaviorism. Logical behaviorists thought that pain was a disposition to pain *behavior*. But as Geach and Chisholm pointed out, what counts as pain behavior depends on the agent's beliefs and desires. Conceptual functionalists avoid this problem by defining each mental state in terms of its contribution to dispositions to behave—and have other mental states.

Functionalism and Physicalism

Theories of the mind prior to functionalism have been concerned both with (1) what there *is* and (2) with what gives each type of mental state its own identity—for example, what pains have in common by virtue of which they are pains. Stretching these terms a bit, we might say that (1) is a matter of ontology and (2) of metaphysics. Here are the ontological claims: dualism told us that there are both mental and physical substances, whereas behaviorism and physicalism are monistic, claiming that there are only physical substances. Here are the metaphysical claims: behaviorism tells us that what pains (for example) have in common by virtue of which they are pains is something behavioral; dualism gave a nonphysical answer to this question, and physicalism gives a physical answer to this question. Turning now to functionalism, it answers the metaphysical question without answering the ontological question. Functionalism tells us that what pains have in common—what makes them pains—is their function; but functionalism does not tell us whether the beings that have pains have any nonphysical parts. This point can be seen in terms of the automaton described above. To be an automaton of the type described, an actual concrete machine need only have states related to one another and to inputs and outputs in the way described. The machine description does not tell us how the machine works or what it is made of, and in particular it does not rule out a machine operated by an immaterial soul, so long as the

soul is willing to operate in the deterministic manner specified in the table. See the papers by Fodor and Putnam in Block 1980.

In thinking about the relation between functionalism and physicalism, it is useful to distinguish two categories of physicalist theses. One version of physicalism competes with functionalism, making a metaphysical claim about the physical nature of mental-state properties or types (and is thus often called ''type'' physicalism). As mentioned above, from one point of view, functionalism shows that type physicalism is false.

However, there are more modest physicalisms whose thrusts are ontological rather than metaphysical. Such physicalistic claims are not at all incompatible with functionalism. Consider, for example, a physicalism that says that every actual thing is made up of entirely of particles of the sort that compose inorganic matter. In this sense of physicalism, most functionalists have been physicalists. Further, functionalism can be modified in a physicalistic direction, for example, by requiring that all properties quantified over in a functional definition be physical properties. Type physicalism is often contrasted with *token* physicalism. (The word ''teeth'' in this sentence has five-letter tokens of three letter types.) Token physicalism says that each pain (for example) is a physical state, but token physicalism allows that there may be nothing physical that all pains share, nothing physical that makes a pain a pain.

It is a peculiarity of the literature on functionalism and physicalism that while some functionalists say functionalism shows physicalism is false (see the papers by Putnam, Fodor, and Block and Fodor in Block 1980, some of which are also in other anthologies), others say functionalism shows physicalism is true (see the papers by Lewis and Armstrong in Block 1980 and Rosenthal 1991). In Lewis's case, the issue is partly terminological. Lewis is a conceptual functionalist about *having pain*. ''Having pain,'' on Lewis's regimentation, could be said to be a rigid designator of a functional property. (A rigid designator names the same thing in each possible world. ''The color of the sky'' is nonrigid, since it names red in worlds in which the sky is red. ''Blue'' is rigid, since it names blue even in worlds in which the sky is red.) ''Pain,'' by contrast, is a nonrigid designator conceptually equivalent to a definite description of the form ''the state with such and such a causal role.'' The referent of this phrase in us, Lewis holds, is a certain brain state, though the referent of this phrase in a robot might be a circuit state, and the referent in an angel would be a nonphysical state. Similarly, ''the winning number'' picks out 17 in one lottery and 596 in another. So Lewis is a functionalist (indeed a conceptual functionalist) about having pain. In terms of the metaphysical issue described above—what do pains have in common by virtue of which they are pains—Lewis is a functionalist, not a physicalist. What my pains and the robot's pains share is a causal role, not anything physical. Just as there is no numerical similarity between 17 and 596 relevant to their being winning numbers, there is no physical similarity between human and Martian pain that makes them pains. And there is no physical similarity of any kind between human pains and angel pains. However, on the

issue of the scientific nature of pain, Lewis is a physicalist. What is common to human and Martian pain in his view is something conceptual, not scientific.

Functionalism and Propositional Attitudes

The discussion of functional characterization given above assumes a psychological theory with a finite number of mental state terms. In the case of monadic states like pain, the sensation of red, and so on, it does seem a theoretical option to simply list the states and their relations to other states, inputs, and outputs. But for a number of reasons, this is not a sensible theoretical option for belief states, desire states, and other propositional-attitude states. For one thing, the list would be too long to be represented without combinatorial methods. Indeed, there is arguably no upper bound on the number of propositions any one of which could in principle be an object of thought. For another thing, there are systematic relations among beliefs—for example, the belief that John loves Mary and the belief that Mary loves John. These belief states represent the same objects as related to each other in converse ways. But a theory of the nature of beliefs can hardly just leave out such an important feature of them. We cannot treat ''believes-that-grass-is-green,'' ''believes-that-grass-is-blue,'' and so forth as unrelated primitive predicates. So we will need a more sophisticated theory, one that involves some sort of combinatorial apparatus. The most promising candidates are those that treat belief as a relation. But a relation to what? There are two distinct issues here. One issue is how to state the functional theory in a detailed way. See Loar 1981 and Schiffer 1987 for a suggestion in terms of a correspondence between the logical relations among sentences and the inferential relations among mental states. A second issue is what types of states could possibly realize the relational propositional-attitude states. Field (1978) and Fodor (in Block 1980) argue that to explain the productivity of propositional-attitude states, there is no alternative to postulating a language of thought, a system of syntactically structured objects in the brain that express the propositions in propositional attitudes. See Stalnaker 1984, chaps. 1–3, for a critique of Field's approach. In later work, Fodor (1987) has stressed the systematicity of propositional attitudes mentioned above. Fodor points out that the beliefs whose contents are systematically related exhibit the following sort of empirical relation: if one is capable of believing that Mary loves John, one is also capable of believing that John loves Mary. Fodor argues that only a language of thought in the brain could explain this fact.

Externalism

The upshot of the famous ''twin-earth'' arguments has been that meaning and content are in part located in the world and in the language community. Functionalists have responded in a variety of ways. One reaction is to think of the inputs and outputs of

a functional theory as *long-arm*—that is, as including the objects that one sees and manipulates. Another reaction is to stick with *short-arm* inputs and outputs that stop at the surfaces of the body, thinking of the intentional contents thereby characterized as *narrow*—supervening on the nonrelational physical properties of the body. There has been no widely recognized account of what narrow content is, nor is there any agreement as to whether there is any burden of proof on the advocates of narrow content to characterize it. See the papers by Burge, Loar, and Stalnaker in Rosenthal 1991; see also Goldman 1993.

Meaning

Functionalism says that understanding the meaning of the word ''momentum'' is a functional state. In one version of the view, the functional state can be seen in terms of the role of the word ''momentum'' itself in thinking, problem solving, planning, and so on. But if understanding the meaning of ''momentum'' is this word's having a certain function, then there is a very close relation between the meaning of a word and its function, and a natural proposal is to regard the close relation as simply identity—that is, the meaning of the word just *is* that function. (c.f. Peacocke 1992.) Thus functionalism about content leads to functionalism about meaning, a theory that purports to tell us the metaphysical nature of meaning. This theory is popular in cognitive science, where in one version it is often known as procedural semantics, as well as in philosophy, where it is often known as conceptual role semantics. The theory has been criticized (along with other versions of functionalism) in Putnam 1988 and Fodor and Lepore 1992.

Holism

Block and Fodor (in Block 1980) noted the ''damn/darn'' problem. Functional theories must make reference to any difference in stimuli or responses that can be mentally significant. The difference between saying ''damn'' and ''darn'' when you stub your toe can, in some circumstances, be mentally significant. So the different functionalized theories appropriate to the two responses will affect the individuation of every state connected to those utterances, and for the same reason, every state connected to those states, and so on. Your pains lead to ''darn,'' mine to ''damn,'' so our pains are functionally different, and likewise our desires to avoid pain, our beliefs that interact with those desires, and so forth. Plausible assumptions lead to the conclusion that two individuals who differ in this way share almost nothing in the way of mental states. The upshot is that the functionalist needs a way of individuating mental states that is less fine-grained than appeal to the whole theory, a molecularist characterization. Even if one is optimistic about solving this problem in the case of pain by finding something

functional that is common to all pains, one cannot assume that success will transfer to beliefs or meanings, for success in the case of meaning and belief may require an analytic-synthetic distinction (Fodor and Lepore 1992).

Qualia

Recall the parity-detecting automaton described at the beginning of this chapter. It could be instantiated by two people, each of whom is in charge of the function specified by a single box. Similarly, the much more complex functional organization of a human mind could ''in principle'' be instantiated by a vast army of people. We would have to think of the army as connected to a robot body, acting as the brain of that body, and the body would be like a person in its reactions to inputs. But would such an army really instantiate a mind? More pointedly, could such an army have pain or the experience of red? If functionalism ascribes minds to things that do not have them, it is too liberal. Lycan (1987) suggests that we include much of human physiology in our theory to be functionalized to avoid liberalism—that is, the theory T in the definition described earlier would be a psychological theory plus a physiological theory. But that makes the opposite problem, chauvinism, worse. The resulting functional description will not apply to intelligent Martians whose physiologies are different from ours. Further, it seems easy to imagine a simple pain-feeling organism that shares little in the way of functional organization with us. The functionalized physiological theory of this organism will be hopelessly different from the corresponding theory of us. Indeed, even if one does not adopt Lycan's tactic, it is not clear how pain could be characterized functionally so as to be common to us and the simple organism. (See my ''Troubles with Functionalism,'' chapter 4, this volume.)

Much of the force of the problems just mentioned derives from attention to phenomenal states like the look of red. Phenomenal properties would seem at least at first glance to be intrinsic to (nonrelational properties of) the states that have them, and thus phenomenal properties seem independent of the relations among states, inputs, and outputs that define functional states. Consider, for example, the fact that lobotomy patients often say that they continue to have pains that feel the same as before, but that the pains do not bother them. If the concept of pain is a functional concept, what these patients say is contradictory or incoherent—but it seems to many of us that it is intelligible. (All the anthologies have papers on this topic. See also Lycan 1987; Shoemaker 1984, chaps. 8, 9, 14, 15; Hill 1991.)

The chauvinism-liberalism problem affects the characterization of inputs and outputs. If we characterize inputs and outputs in a way appropriate to our bodies, we chauvinistically exclude creatures whose interface with the world is very different from ours, for example, creatures whose limbs end in wheels, or turning to a bigger difference, gaseous creatures who can manipulate and sense gases but for whom only the

topologies of solids and liquids matter. The obvious alternative of characterizing inputs and outputs themselves functionally would appear to yield an abstract structure that might be satisfied by, for instance, the economy of Bolivia under manipulation by a wealthy eccentric, and would thus fall to the opposite problem of liberalism.

It is tempting to respond to the chauvinism problem by supposing that the same functional theory that applies to me also applies to the creatures with wheels. If they thought they had feet, they would try to act like us, and if we thought we had wheels, we would try to act like them. But notice that the functional definitions have to have some specifications of output organs in them. To be neutral among all the types of bodies that sentient beings could have would just be to adopt the liberal alternative of specifying the inputs and outputs themselves functionally. Some suppose that the problem can be handled by conditional outputs. For example, wanting to get the ball to the end of the field could be defined in part by the tendency to kick it if one has limbs of a certain sort, push it if one has limbs of another sort, etc. But it is not clear that the ''etc.'' could ever be filled in, since it would require enumerating and physically describing every kind of output organ an intelligent being could have. Further, the result of a certain desire on one's limbs depends on how they are wired up as well as their physical shape. In the context of the ''wrong'' wiring, a desire to get the ball to the end of the field would result in the ball stuck in one's mouth rather than propelled down the field. But this point makes it look as if the problem will require going far beyond anything that could be said to be implicit in common sense. (See Braddon-Mitchell and Jackson 1995, for an opposing view.)

Teleology

Many philosophers (see the papers by Lycan and Sober in Lycan 1990, as well as Lycan 1987) propose that we avoid liberalism by characterizing functional roles teleologically. We exclude the armies and economies mentioned because their states are not *for* the right things. A major problem for this point of view is the lack of an acceptable teleological account. Accounts based on evolution smack up against the swamp-grandparents problem. Suppose you find out that your grandparents were formed from particles from the swamp that came together by chance. So, as it happens, you do not have any evolutionary history to speak of. If evolutionary accounts of the teleology underpinnings of content are right, your states do not have any content. A theory with such a consequence should be rejected.

Causation

Functionalism dictates that mental properties are second-order properties, properties that consist in having other properties that have certain relations to one another. But

there is at least a prima facie problem about how such second-order properties could be causal and explanatory in a way appropriate to the mental. Consider, for example, provocativeness, the second-order property that consists in having some first-order property (say redness) that causes bulls to be angry. The cape's redness provokes the bull, but does the cape's provocativeness provoke the bull? The cape's provocativeness might provoke an animal protection society, but isn't the bull too stupid to be provoked by it? See Block 1990.

Functionalism continues to be a lively and fluid point of view. Positive developments in recent years include enhanced prospects for conceptual functionalism and the articulation of the teleological point of view. Critical developments include problems with causality and holism, and continuing controversy over chauvinism and liberalism.

Note

This is a somewhat revised version of an entry in *The Encyclopedia of Philosophy*, supplement (New York: Macmillan, 1997).

References

Beakley, Brian, and Ludlow, Peter, eds. 1992. *Philosophy of Mind: Classical Problems/Contemporary Issues*. Cambridge, MA: MIT Press.

Block, Ned, ed. 1980. *Readings in the Philosophy of Psychology*. 2 vols. Cambridge, MA: Harvard University Press.

Block, Ned. 1990. Can the mind change the world? In G. Boolos, ed., *Meaning and Method: Essays in Honor of Hilary Putnam*, 137–170. Cambridge: Cambridge University Press.

Braddon-Mitchell, David, and Jackson, Frank. 1995. *Philosophy of Mind and Cognition*. Oxford: Blackwell.

Chisholm, Roderick. 1957. *Perceiving*, chap. 11. Ithaca, NY: Cornell University Press.

Field, Hartry. 1978. Mental representation. *Erkentniss* 13: 9–61.

Fodor, Jerry. 1987. *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, Jerry, and Lepore, Ernest. 1992. *Holism*. Oxford: Blackwell.

Geach, Peter. 1971. *Mental Acts*. Chicago: St. Augustine's Press.

Goldman, Alvin. 1993. *Readings in Philosophy and Cognitive Science*. Cambridge, MA: MIT Press.

Hill, Christopher S. 1991. *Sensations*. Cambridge: Cambridge University Press.

Lewis, David. 1995. Reduction of mind. In S. Guttenplan, ed., *A Companion to the Philosophy of Mind*. Blackwell: Oxford.

Loar, Brian. 1981. *Mind and Meaning*. Cambridge: Cambridge University Press.

Lycan, William G. 1987. *Consciousness*. Cambridge, MA: MIT Press.

Lycan, William G., ed. 1990. *Mind and Cognition*. Oxford: Blackwell.

Peacocke, C. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.

Putnam, Hilary. 1988. *Representation and Reality*. Cambridge, MA: MIT Press.

Rosenthal, David, ed. 1991. *The Nature of Mind*. Oxford: Oxford University Press.

Schiffer, Stephen. 1987. *Remnants of Meaning*. Cambridge, MA: MIT Press.

Shoemaker, Sydney. 1984. *Identity, Cause, and Mind*. Ithaca, NY: Cornell University Press.

Stalnaker, Robert C. 1984. *Inquiry*. Cambridge, MA: MIT Press.