Philosophical Perspectives, 3
Philosophy of Mind and Action Theory, 1989

ON A CAUSAL THEORY OF CONTENT

Lynne Rudder Baker
Middlebury College

The project of explaining intentional phenomena in terms of nonintentional phenomena has become a central task in the philosophy of mind.[1] Since intentional phenomena like believing, desiring, intending have *content* essentially, the project is one of showing how semantic properties like content can be reconciled with nonsemantic properties like cause. As Jerry A. Fodor put it,

> The worry about representation is above all that the semantic (and/or the intentional) will prove permanently recalcitrant to integration in the natural order; for example that the semantic/intentional properties of things will fail to supervene upon their physical properties.[2]

What is wanted is "a *naturalized* theory of meaning; a theory that articulates, in nonsemantic and nonintentional terms, sufficient conditions for one bit of the world to be *about* (to express, represent, or be true of) another bit." (p. 98)

My aim here is to examine Fodor's own response to the worry about representation, and to show that his Causal Theory of Content is not up to the task of naturalizing intentionality. Since Fodor's recent work is the most developed and most promising attempt to give an explicit solution to the "naturalization" problem, if I am right about that attempt, then very little headway has been made toward naturalizing intentionality in the intended sense.

Psychosemantics

Fodor offers a Representational Theory of Mind, according to which "to believe that *P* is to bear a certain relation to a token of a symbol which means that *P*."(p. 135) The immediate and pressing question is this: what makes a token mean that *p*? What makes a given token a 'red'-thought or a 'water'-thought?[3] More generally, how is the nonlogical, primitive vocabulary of the language of thought (Mentalese) to be interpreted? The task of psychosemantics is to specify sufficient conditions, in nonsemantical and nonintentional terms, for a given mental token to represent *A*.

Fodor begins by formulating a crude (and obviously false) view that he proceeds to refine: A symbol expresses a property, and its tokens denote the property, "if it's nomologically necessary that *all* and *only* instances of the property cause tokenings of the symbol."(p. 100)

Both the "only" and "all" clauses are then qualified. The "only" clause is modified to accommodate misrepresentation, and a necessary condition for "wildness" underwrites the modification: Only A's cause non-wild 'A'-tokenings. The "all" clause is restricted in application to observation terms. What is claimed for the resulting theory, the Slightly Less Crude Causal Theory of Content, is that

> it does what metaphysical skeptics about intentionality doubt *can* be done: it provides a sufficient condition for one part of the world to be semantically related to another part (specifically, for a certain mental representation to express a certain property); it does so in nonintentional, nonsemantical, nonteleological, and in general, non-question-begging vocabulary; and it's reasonably plausible....(p. 98)

Although I am no metaphysical skeptic about intentionality, I do not believe that the Slightly Less Crude Causal Theory of Content accomplishes what is claimed for it. In particular, I think that the "only" clause fails, because the account of misrepresentation or wildness actually applies to nothing; and I think that the "all" clause fails, because no psychophysically optimal conditions can be antecedently specified that guarantee that red instantiations produce 'red'-tokenings. I shall argue for these claims in the next two sections.

Misrepresentation

The tenet of the Crude Causal Theory that only A's cause 'A'-tokens leads straight to what Fodor calls 'the disjunction problem:' Suppose that horses reliably cause tokens of a certain type, but that sometimes cows cause tokens of that type. Although what we would like to say on such occasions is that a cow is misrepresented as a horse, what the "only" clause of the crude theory forces us to say is something else, something that eliminates the possibility of error: since cows as well as horses are sufficient to produce tokens of the given type, we must interpret the type as expressing not the property *horse*, but rather the disjunctive property *horse or cow*.[4] Since this has the consequence that misrepresentation of a cow as a horse is impossible, the "only" clause of the Crude Causal Theory must be modified.

To distinguish error from genuinely disjunctive properties, Fodor draws upon the observation that falsehoods are ontologically dependent upon truths. An instance of a B can cause an 'A'-tokening only when there is independently a semantic relation between A's and 'A'-tokenings; and the fact that a B causes an 'A'-tokening depends on the fact that A's cause 'A'-tokenings, but the fact that A's cause 'A'-tokenings does not depend on the fact that B's cause 'A'-tokenings. For example, a cow-caused token is a misrepresentation of a horse only if the fact that a cow causes the token depends on the fact that horses cause tokens of that type, but the fact that horses cause tokens of that type does not depend on the fact that a cow ever causes tokens of that type.

Thus, the causal connection between B's and 'A'-tokenings is asymmetrically dependent on the causal connection between A's and 'A'-tokenings. This seems to yield a necessary condition for misrepresentation, or wildness: "B-caused 'A'-tokenings are wild only if they are asymmetrically dependent upon non-B-caused 'A'-tokenings."(p. 108) Since the wild B-caused token would not have been an 'A'-token without an independent relation between A's and tokens of that type, a case of error without asymmetric dependence on truth would be a counterexample to this account of misrepresentation.[5]

Let us consider this account in light of a particular case. Suppose that, although there are many ordinary cats around, a certain person, *S*, learns a particular Mentalese symbol solely from artifacts (say, Putnam's robot-cats) that impinge on sensory surfaces in exactly the same way as cats.[6] Now (for the first time) *S* sees a real cat, which

happens to be chasing a robot-cat. How should Fodor interpret the cat-caused token? His theory should apply to such a case, for *S* is in a world with the same fundamental physical laws as our world, and in a world in which reference is possible; the only difference between the imagined world and the actual world is a slight difference in environment—just the kind of difference to which reference is supposed to be responsive.

There seem to be three possibilities for interpreting the cat-caused token. Either (i) it is a correct representation of a cat (and the robot-caused tokens are misrepresentations of cats); or (ii) it is a misrepresentation of a robot-cat (and the robot-caused tokens are correct representations of robot-cats); or (iii) it is a correct representation of the disjunctive property *cat-or-robot-cat*.

(i) Suppose that the cat-caused token is a correct representation of a cat and the robot-caused tokens are misrepresentations of cats. Although this construal may seem implausible, Fodor explicitly allows it when he assumes that one can "learn 'horse' entirely from noninstances."(p. 109) So, for the moment, put aside qualms about what, on a Causal Theory, would make 'horse' the correct interpretation of a mental symbol whose tokening has never been caused by a horse and suppose that *S* has learned 'cat' entirely from noninstances. In that case, if Fodor's account is correct, there must be asymmetric dependence of the content of robot-caused tokens on the content of *S*'s single cat-caused token. But as the story was told, the requisite dependence is missing; if there is any asymmetric dependence, it goes the other way. *S*'s present disposition to apply 'cat' to a real cat depends upon her corresponding current disposition to apply it to robot-cats.[7] So, Fodor can not take the cat-caused token to be a representation of a cat.[8]

(ii) On the other hand, suppose that we interpret the robot-caused tokens as 'robot-cat'-tokens, and then say that the lone cat-caused token is a misrepresentation of robot-cats as cats. Not only does this move fly in the face of Fodor's assumption that one can learn a mental symbol entirely from noninstances, but also it opens Fodor up to the criticism that he leveled against Dretske (p. 104): this move ignores relevant counterfactuals. Since encounters with real cats, which are plentiful, would have produced the same type of tokens that were in fact produced by robot-cats, there seems to be no asymmetry whatever between tokens caused by cats and tokens caused by robot-

cats, and hence no misrepresentation at all, on Fodor's account. The correlation that was established is "(not a nomological dependence of 'A's on A's but) a nomological dependence of 'A's on (A v B)s."(p. 104) In the case at hand, the correlation is between tokens of a certain type and (cats or robot-cats). It is simply an accident that the actual causes of *S*'s early representations were all robot-cats; representations of the same type would have been caused by real cats. So, consideration of counterfactuals suggests that we can not take the cat-caused token to be a misrepresentation of a robot-cat.

(iii) The symmetry between the cat-caused token and the robot-caused tokens just rekindles the disjunction problem. If cats and robot-cats are both sufficient for the causation of tokens of a certain type and if "symbols express the properties whose instantiations reliably cause them," then what tokens of that type must express is not the property *cat* (or the property *robot-cat*) but rather the disjunctive property *cat-or-robot-cat*. In that case, both the cat-caused and the robot-caused tokens are veridical after all—even when *S*, on subsequently discovering the difference between cats and robot-cats, exclaims, "I mistook that robot for a cat!" Fodor's account seems to preclude saying that *S* made an error, describable as mistaking a robot for a cat. We would have to say that her mistake was to think that she had made a mistake, and try (perhaps without success) to find some way to make sense of her "second-order" mistake. Since the same story could be told about any other mental symbol (at least any other mental symbol for nonobservables), this leaves us with no general account of misrepresentation.

The only way I see to reply is to confine the account of error to observation terms. In that case, no interpretation of the story about cats and robot-cats would impugn Fodor's account of error, because the account of error would not be taken to apply to natural kind terms anyway. However, this reply does not save the account of error in terms of asymmetric dependence.

First, the fact that Fodor develops his account of error with respect to 'horse' suggests that he wants it to apply more broadly than merely to cases concerning observation terms. Unless Fodor can find a way to develop an account of error for terms like 'cat' or 'horse,' it would seem that errors about cats and horses would have to be accounted for in terms of errors about more observable shapes, colors and so on. But in the cat/robot-cat case, there are no errors concerning such observables. So, confining the account of error to observation terms

would still just leave us with no account of error for horses or cows.

Moreover, confining the account of error to observation terms—even if Fodor turned out to be right about observation terms—would have the effect of making the account of error apply to nothing at all. For Fodor builds an account of error (one not based on asymmetrical dependence) into his account of the semantics of observation terms. Fodor invokes psychophysics, which specifies in nonsemantic and nonintentional terms circumstances—e.g., the lighting is such and such, the visual system intact, and so on—in which observation terms are tokened. If *S* is in psychophysically optimal circumstances and *red* is instantiated, then by psychophysical law, *S* can not avoid producing a 'red'-token. The only source of error for observation terms is less-than-optimal circumstances; as Fodor describes psychophysics and its application to observation terms, we need no further account of error in terms of asymmetric dependence for observation terms.

Therefore, Fodor's account of misrepresentation solves no problems: For nonobservation terms, it leaves the disjunction problem in place and fails to provide a necessary condition for error, and for observation terms, it is otiose. But without an account of misrepresentation or wildness that has appropriate application, the change from the unmodified "only" clause of the Crude Causal Theory to the modified "only" clause of the Slightly Less Crude Causal Theory of Content has no effect. In the absence of a successful account of wildness, "Only A's cause nonwild 'A'-tokenings," or "All non-A-caused 'A'-tokenings are wild" is no better than "Only A's cause 'A'-tokenings," and the account remains inadequate.

Let me close this discussion of misrepresentation with a conjecture. Not only is the particular account of error in terms of an asymmetrical dependence on truth inadequate, but I believe that no such account, under the constraints imposed by the project of naturalizing intentionality, can succeed. My doubts stem from my suspicion that any such account entails a kind of type/type identity, which I think false.

To see how there can be error, Fodor says, "we need a difference between A-caused 'A'-tokenings and B-caused 'A'-tokenings that can be expressed in terms of nonintentional and nonsemantic properties of causal relations."[9] But if 'A'-tokenings are taken to represent A's, the question arises: What nonsemantic, nonintentional conditions make 'an 'A'-token' even a candidate as a description of the

token caused by B? I suspect that the answer to this question leads straight into the arms of type/type identity.

To see why, consider Fodor's own description of the problem of error: "I see a cow which, stupidly, I misidentify. I take it, say, to be a horse. So taking it causes me to effect the tokening of a symbol; viz., I say 'horse.'"(p. 107) What does the error consist in? Fodor's answer is that there is "independently a semantic relation between 'horse'-tokenings and horses," and on this occasion a cow caused a 'horse'-tokening.

What nonintentional, nonsemantic conditions determine that the token caused on this occasion by a cow is a 'horse'-token? Whatever makes the Mentalese token a 'horse'-token can in no way depend on the English word 'horse' since a main point of psychosemantics is to show how mental tokens can represent without presupposing a public language. So if Fodor reports his token by saying in English, 'There's a horse,' that report is not constitutive of the error. Such a report is a semantic feature of the case, irrelevant under the constraints of psychosemantics. Furthermore, the fact that on this occasion, the cow looked like a horse to Fodor is an intentional feature and also is irrelevant.

So, in virtue of what is this cow-caused token a 'horse'-token? The only answer that I can think of in terms of nonsemantic and nonintentional conditions is this: (1) the causal relation previously established between horses and 'horse'-tokens was in fact a causal relation between horses and mental tokens of a nonsemantic type T; and (2) mental tokens of nonsemantic type T are 'horse'-tokens in virtue of this causal relation; and finally (3) the cow on this occasion caused a token of nonsemantic type T.

But if (1)—(3) give an accurate account of why the cow-caused token is a 'horse'-token, then Fodor's description of error commits him to a kind of type/type identity of the semantic and the nonsemantic. (1)—(3) entail type/type identity, because they entail that all tokens of a single semantic type (at least within an individual) are of a single nonsemantic type; what makes a token a 'horse'-token is that it is of a particular nonsemantic type caused by horses. But to suppose that all tokens that represent horses are of a single nonsemantic type is wildly implausible.[10]

Thus, we are left with no satisfactory account of misrepresentation. The notion of asymmetrical dependence of misrepresentation on accurate representation turned out to solve no problems, and,

if my conjecture is correct, any account conforming to the constraints of psychosemantics will be committed to an implausible type/type identity theory. Therefore, I think that no account of error in terms of nonintentional and nonsemantic conditions is likely to succeed.

Local Color

The "all" clause of the Slightly Less Crude Causal Theory of Content, even when restricted to terms like 'red,' is no less flawed than the "only" clause.

For a certain class of mental symbols—so-called observation terms—Fodor claims that psychophysics, coupled with a Causal Theory of Content, solves the naturalization problem by providing "a plausible sufficient condition for certain symbols to express certain properties: viz., that tokenings of those symbols are connected to instantiations of the properties by psychophysical law."(p. 113)

Even if this is correct, Fodor points out, there is no obvious way to extend the naturalized account to nonobservation terms generally. Yet we need to be able to extend it "on pain of having the metaphysical worry that—excepting psychophysical concepts—*we have no idea at all* what a naturalized semantics would be like for the nonlogical vocabulary of Mentalese."(p. 118) I shall say little about Fodor's approach to extending naturalization to nonobservation terms; for I think that the difficulties arise right at the beginning.

In this section, I shall be concerned with the claim that the Causal Theory of Content, together with psychophysics, approximates a "complete solution to the naturalization problem" for a mind whose nonlogical vocabulary consists exclusively of observation terms like 'red.' In the next section, I shall consider the prospects for any project that relies on a comprehensive, context-free distinction between observation and nonobservation terms.

What psychosemantics needs are nonintentionally and nonsemantically specifiable sufficient conditions under which instantiations of *red* produce tokens that denote RED (or, equivalently, as I am using these terms, 'red'-tokens).

> All that matters is that there are concepts (Mentalese terms) whose tokenings are determined by psychophysical law; and that these concepts are plausibly viewed as expressing the properties upon which their tokening is thus lawfully

> contingent; and that the psychophysical conditions for the tokenings of these concepts can be expressed in nonintentional and nonsemantical vocabulary.(p. 113-14)

It is not required, Fodor emphasizes, "that we view psychophysics as enunciating sufficient conditions for the fixation of *belief*"; the purpose will be served "if psychophysics enunciates sufficient conditions for the fixation of *appearances*."(p. 114) What one believes depends in part on the way things appear and in part on one's cognitive background.

> [Psychophysics] can't guarantee that you'll *believe* 'red there,' only that 'red there' will occur to you. But a guaranteed correlation between instances of red and tokenings of 'red there' in the occurs-to-me box will do perfectly nicely for the purposes of semantic naturalization; all semantics wants is *some* sort of nomologically sufficient conditions for instances of *red* to cause psychologically active tokenings of 'red.'(p. 114)

The claim, then, is that psychophysics specifies conditions in which instantiations of *red* appear red, or alternatively, conditions in which instantiations of *red* produce a token of 'red there.'

There is, I believe, a fundamental incoherence in this view, which elsewhere I have tried to bring out by thought experiments about molecular duplicates and their production of 'red'-tokens.[11] Here, I want to come at it from a different perspective—one more epistemic than semantic.

The basic epistemic difficulty, one that almost exactly parallels the basic semantic difficulty, can be illustrated by an example. So, to begin, consider three events on three possible worlds, which differ in local environmental conditions to be described shortly. There are no differences in physical laws among the worlds, only environmental differences that scientific laws should be able to accommodate. Let *S* be a sincere and competent English-speaking Earthian.

> Event 1: On Earth, in conditions psychophysically optimal on Earth, specified in terms of "wavelengths, candlepowers, retinal irradiations, and the like" *S* is placed in an empty room, with solid red walls.
> Event 2: On Mars, in conditions psychophysically optimal on

> Mars, specified in terms of "wavelengths, candlepowers, retinal irradiations, and the like," *S* is placed in an empty room with solid green walls.
> Event 3: On Venus, in conditions psychophysically optimal on Venus, specified in terms of "wavelengths, candlepowers, retinal irradiations, and the like," *S* is placed in an empty room with solid red walls.

Suppose that on Mars, atmospheric conditions transform light waves of the lengths that green objects on Earth reflect into waves of the lengths that red objects on Earth reflect. For example, if a green object is transported to Mars without any change in the object (such as painting it red), light reflected by that object on Mars has the same retinal effect as light reflected by red objects on Earth. Physical measurements reveal that light at the surface of the object on Mars is of the same wavelengths as light reflected by green objects on Earth, but that light 1 centimeter away from the object on Mars is of the same wavelengths as light reflected by red objects on Earth. Otherwise, Mars and Earth are similar. So, when *S* is placed in the green room on Mars, her visual system is affected in the way that it was when she was placed in the red room on Earth.

On Venus, by contrast, atmospheric conditions transform light waves of the lengths that red objects on Earth reflect into waves of the lengths that green objects on Earth reflect. For example, if a red object is transported to Venus without any change in the object (such as painting it green), light reflected by that object on Venus has the same retinal effect as light reflected by green objects on Earth. Physical measurements reveal that light at the surface of the object on Venus is of the same wavelengths as light reflected by red objects on Earth, but that light 1 centimeter away from the object on Venus is of the same wavelengths as light reflected by green objects on Earth. Otherwise, Venus and Earth are similar. So, when *S* is placed in the red room on Venus, her visual system is affected in the way that it would be if she were placed in a green room on Earth.

Now, ask: Do the walls appear the same to *S* on Earth and on Mars? On Earth and on Venus? In each case, one answer denies the Causal Theory of Content, and the other answer detaches appearances from internal physical states. I shall defer comment on an objection about optimal conditions until later; it will be easier to defuse the objection in light of attempts to answer these questions.

First, compare Earth and Mars. Do the walls appear the same to *S* on both? Suppose that the answer is yes. Then we would have to say that different causes (red instantiations and green instantiations) in optimal circumstances can produce the same appearances. In that case, we have a disjunction problem. For the token produced should not be interpreted as a 'red'-token, but as a 'red or green'-token. Even worse, in various other worlds, atmospheric conditions could transform wavelengths for blue, green, magenta or even wavelengths for Middle C into wavelengths produced by red things on Earth. Thus, there would be no limit on the different property instantiations sufficient in optimal conditions for the production of a token of a given physical type. However, to say that the interpretation of the token is as a 'red or green or magenta...or Middle C'-token is simply to say that we have no interpretation at all. The reappearance of the disjunction problem thus disables the Less Crude Casual Theory on this alternative.

So, suppose that the walls do not appear the same to *S* on both Earth and Mars. In the example, *S* is in the same internal physical state on Earth and Mars; her sensory stimulation is the same on both. Therefore, on this alternative a single internal state can subserve two types of appearances. Not only is this a violation of mind/brain supervenience, but also, since there are endless combinations of cause and atmospheric conditions in different worlds that would have the same sensory effects, a single internal state can subserve any appearance, on this alternative. Thus, neither answer to the question of whether the walls appear the same to *S* on Earth and Mars seems satisfactory.

Do the walls appear the same to *S* on Earth and Venus? Suppose that the answer is yes. Then since *S*'s sensory stimulation on Venus is what it would be if produced in optimal conditions by green things on Earth, this answer, like the negative answer above, has the effect of detaching appearances from internal physical state. For on this alternative, different internal states subserve the same type of appearance. Again, endless combinations of cause and atmospheric conditions in various possible worlds would produce a single type of internal state. On this alternative, then, any internal state can subserve a given appearance.

So, suppose that the walls do not appear the same to *S* on Earth and on Venus. But since *S* is in optimal conditions on both, and red is instantiated on both, the putative psychophysical law must be false

if the walls do not appear the same to *S* on both. This alternative thus becomes a straightforward counterexample to the Causal Theory of Content. Thus, neither answer to the question of whether the walls appear the same to *S* on Earth and Venus seems satisfactory. There simply seem to be no coherent answers to the questions of whether or not the walls appear the same to *S* on Earth and on Mars, on the one hand, and on Earth and on Venus, on the other.

There may seem to be an easy reply in terms of optimal conditions: Since my discussion relies on a difference in atmospheric conditions on Earth, Mars and Venus, perhaps we should make reference to atmospheric conditions in the specification of optimal conditions. There are two ways to incorporate atmospheric conditions: (i) we may continue to take 'optimal conditions' nonrigidly to refer to whatever conditions are optimal in a given world, or (ii) we may take 'optimal conditions' rigidly to refer to the conditions optimal on Earth.

(i) If we continue to relativize optimal conditions to world, then adding reference to atmospheric conditions makes no difference to the result of the discussion of Events 1, 2 and 3; for even though on this alternative, there is no single set of optimal conditions across possible worlds, for each Event, *S* is in optimal conditions in the world in which he is located. That is, on the nonrigid construal, since *S* is in atmospheric conditions optimal in each world, the objection that optimal conditions must include atmospheric conditions is just irrelevant to my point.

(ii) So, consider the rigid construal. Suppose that the only psychophysically relevant conditions are those that are optimal on Earth. I think that this construal is unsatisfactory for two reasons. First, it would entail, implausibly, that psychophysically optimal conditions never obtain on Mars. In any case, we need not go to Mars to see the implausibility of taking conditions that never obtain to be optimal. Suppose that, as a result of pollution, Earth's atmosphere changed in such a way that objects that formerly reflected light of green wavelengths then reflected light of red wavelengths. Suppose that 50,000 years later, scientists took the prevailing conditions to be optimal and formulated (what they took to be) psychophysical laws that were as well-confirmed as psychophysical laws are today. But if optimal conditions are those that are now optimal on Earth, as the rigid construal supposes, then psychophysics on future Earth would be impossible; for optimal conditions on future Earth, on the rigid construal, would simply never obtain. To suppose that there can be no psychophysics in such environmentally-altered cir-

cumstances seems to me a kind of planetary and temporal chauvinism.[12]

Second, the rigid construal of 'optimal conditions' in effect gives up the Causal Theory by reducing it to the role of a reference fixer. Here is the reason: The Causal Theory of Content is supposed to be a general theory of 'aboutness,' a phenomenon not confined to Earth. So, even if we take optimal conditions rigidly to be those optimal on Earth (including atmosphere), we still must be able to make comparisons between Earth and Mars. On this construal, do the walls appear the same to *S* on Earth and on Mars? If not, then (again) appearances get detached from internal states. So assume that the walls on Earth and on Mars *do* appear the same to *S*. The reason would be this: *S* is in the same internal sensory state on Mars that she would be on Earth when her mental token on Earth is a 'red'-token. But if *S* has the same appearances on Earth and on Mars on this rigid construal of 'optimal conditions,' the theoretical work would be done by type-type identity: Without type-type identity, there would be no grounds for supposing that the green instantiation on Mars produced a red appearance. What makes an appearance a 'red'-appearance, on this alternative, is its physical type, not its causal history.

Compare: we can fix the referent of 'water' by properties like being wet, being drinkable, and so on. But what determines that a particular liquid is water are not these properties at all, but its being a sample of a certain chemical type, H_2O, as opposed to a sample of the equally wet and drinkable XYZ. Similarly, on the alternative under consideration, what determines that an appearance is red is its being of a certain physical type. For example, we say that the grey-caused token on Mars is a 'red'-token only because of physical similarities between it and 'red'-tokens on Earth. Invocation of causal histories of tokens of that physical type in optimal circumstances merely serves to pick out the relevant physical type. Since content here is determined by physical type, not causal history, this alternative just abandons the Causal Theory of Content.

Thus, the objection concerning optimal conditions does not affect my argument on the nonrigid construal, and leads away from a Causal Theory of Content on the rigid construal. From the point of view of the Book of Nature, the conditions that we select as optimal seem to have no special status anyway. Insofar as a psychophysical law is written in the Book of Nature, its antecedent is an infinitely long disjunction of causes and conditions, one disjunct of which is of par-

ticular interest to us and is singled out as specifying conditions that are optimal. There is the same lawlike connection between red walls, one kind of light, and production of tokens of a certain type, on the one hand, as there is between green walls, a different kind of light, and production of tokens of the same type, on the other. Nomologically speaking, these are on a par. Conditions are optimal relative to our interests; optimality is not given by nature.

Why is the Causal Theory of Content unable to deliver answers to questions about the way the walls appear to *S*? The reason, I think, is that the Causal Theory implicitly employs two criteria for identity of appearances:

(A1) Same environmental causes (in optimal circumstances) --> same appearances.
(A2) Same sensory effects (in optimal circumstances) --> same appearances.

(A1) and (A2) could peacefully co-exist if any two appearances that were counted as the same according to one were counted as the same according to the other. But in trying to answer the questions about how the walls in Events 1, 2, and 3 appeared to *S*, we have seen that (A1) and (A2) come apart. Since their implicit conjunction leads to incoherence, perhaps we should just abandon one of them.

To give up (A1) would be to give up the Causal Theory of Content. For if we gave up (A1) in favor of (A2), we should say that the walls on Earth and on Mars appeared the same, because they have the same sensory effects. But different color properties were instantiated (even in optimal conditions) on Earth and on Mars. Therefore, on this alternative, types of appearance are not even correlated with types of causes in optimal circumstances, and *red* instantiations/*green* instantiations do not determine different appearances (red appearances/green appearances).

However, to give up (A2) would also be unpromising from the point of view of the naturalization problem. If we gave up (A2) in favor of (A1), we should say that the walls on Earth and on Venus appeared the same to *S*. But consider *S*'s point of view. This second alternative requires us either to disregard *S*'s point of view altogether, or to take what a sincere and competent speaker says to be *no* evidence of the way things appear to her; in either case, it cuts thought totally off from public language.

To see this, suppose that we asked *S* to compare how the walls appeared to her on Earth and on Venus. She would reply that on Earth the room appeared red to her. Since on Venus, the red room impinged on *S*'s senses in the same way as a green room would on Earth, she would say that on Venus, the walls appeared green to her. Even if informed of the different color property instantiations and the differences in atmosphere, when asked how the rooms in each case appeared, she would still respond that they looked different on Earth and on Venus—just as an expert on refraction would say that the stick in water still *looked* crooked. However, on the option of giving up (A2) and retaining (A1), it follows that 'red there' in Mentalese occurred to *S* on Venus, despite the fact that she is a competent English speaker who insists that 'green there' occurred to her.

It is important to see that the difficulty here has nothing to do with any assumptions about incorrigibility. *S* is not exactly making a mistake about the way that she is being appeared to. Rather, her thought-token receives incompatible interpretations: 'red there' in Mentalese and 'green there' in English. Perhaps, alternatively, we could say that she is thinking two incompatible thought-tokens at once. Either way, there is an incoherence between Mentalese and English; different properties are being represented simultaneously. Unless this incoherence can be overcome (and I do not see how it can be), on the Causal Theory of Content, we must disregard all the subject's testimony.[13]

A corollary of this (still assuming that we give up (A2) in favor of (A1)) is that we all lack verbal access to what occurs to us, to the ways that things appear to us; our reports and any thoughts that we can put into words are no evidence at all as to the actual character of the way that things appear. Such a view would make the relations between spoken language and thought a total mystery. How could language be the expression of thought, as Fodor says it is, if the theory of thinking forces us to say that things appear red to a sincere and competent speaker of English who says that they appear not red, but green?

And to hold on to the Causal Theory of Content at the cost of saying that, his denials notwithstanding, the walls appear red to *S* vitiates Fodor's entire discussion in which he replaces Mentalese with English as the language of thought. In his attempt to solve the disjunction problem, Fodor says that he sets it up, "as it happens, for a token

of English rather than a token of Mentalese; but none of the following turns on that."(P107) But if we lack verbal access to our Mentalese tokens—if 'red there' can occur to us when we would swear that 'green there' occurs to us—a lot turns on the substitution of English tokens for Mentalese tokens. The shift from Mentalese to English is thus far from innocent.[14]

Let me note two further consequences of giving up (A2): (1) any internal perceptual state could subserve any appearance whatsoever. (See initial discussion of Events 1, 2, and 3.) This makes internal physical state almost irrelevant to 'red'-tokenings.[15] (2) It seems that no contentful state ever supervenes on the brain. If, as Fodor argues (p. 33-44), causal powers do supervene on the brain, we seem to have the further consequence that causal powers and content swing free of each other.[16]

Before making a final determination that the conjunction of (A1) and (A2) leads to incoherence, let us consider a re-description of the cases. If we take optimal conditions to be those in which things appear the way that they are, we could re-describe the Event on Mars (and similarly the Event on Venus) in one of three ways: (a) We could take the walls on Mars to be green and *S* to be in optimal conditions, and then conclude that *S*'s appearance was green. But this interpretation violates mind/brain supervenience by detaching internal state from appearance, and disregards *S*'s point of view altogether—in the same way that we saw when we considered giving up (A2) below.

(b) We could take the walls on Mars to be green and *S*'s appearance to be red, and then conclude that the conditions are not optimal. However, this suggestion allows conditions that *never* obtain to be optimal. Such a result would be unsatisfactory for reasons given in discussing the rigid construal of 'optimal conditions.'

(c) We could take *S* to be in optimal conditions on Mars and *S* to have a red appearance, and then conclude that the walls on Mars are red. Since light measured at the surface of the walls on Mars reflected wavelengths of green, this suggestion would detach the color of the walls from their reflective properties. Perhaps, then, one would hold that colors of objects are determined by their reflective properties *together with* atmospheric conditions. But suppose that the Martian atmosphere is intermittently polluted in such a way that colors look different on some days than on others (as actually happens on Earth). In that case, the current suggestion, which takes atmosphere as a determinant of actual color of objects rather than as

a condition that may or may not be optimal for viewing would force us to say that the actual colors of objects, not just how the objects look, change from day to day. I do not think that we understand color in this way, nor do I think that we would if Earth's atmosphere were like Mars' supposed atmosphere.[17]

It seems, then, that although the conjunction of (A1) and (A2) leads to incoherence, the Causal Theory of Content can not dispense with either one.[18] Since the problems are perfectly general, it would be wrong to suppose that since contexts often do affect brain states, the difficulties raised here may be confined to "funny cases," as Fodor suggests (p. 159, 10n); for they arise with respect to 'red' or any other so-called observation term.

For these reasons, I do not believe that the Causal Theory of Content, coupled with psychophysics, has begun to solve the naturalization problem for observation terms, as Fodor formulates it. Therefore, it seems that neither the modified "all" clause nor the modified "only" clause of the Slightly Less Crude Causal Theory of Content can deliver what is claimed for it.

Saving Appearances?

Is a comprehensive, context-free distinction between observation and nonobservation terms a viable basis for any semantic theory? I am doubtful. For it is not even clear that 'red' is an instance of a semantically special type of term, a class that is supposed to include other terms traditionally linked to sensory experience and to exclude terms like 'water' whose interpretations are less directly accessible to the senses. On the one hand, comparison of 'red' with observationally equivalent observation terms 'magenta' and 'Middle C' shows that what is to distinguish 'red' as an observation term does not hold in the case of the other terms. On the other hand, comparison of 'red' with a nonobservation term like 'water' reveals an important semantic symmetry. Thus, the claim that there is a special class of observation terms and "the idea that the semantics of observation terms is somehow at the core of the theory of meaning"(p. 114) are not borne out.

What is supposed to distinguish terms like 'red' as observation terms?

> [W]hat makes RED special—what makes it a 'psychophysical concept' within the meaning of the act—is that the difference between merely seeing something red and succeeding in seeing it *as* red vanishes when the observer's point of view is psychophysically optimal. You can't—or so I claim—see something red under psychophysically optimal viewing conditions and *not* see it as red.

The Venus case has already given reason to be skeptical of this: under psychophysically optimal viewing conditions on Venus, our subject saw red things, but did not see them *as* red. But I don't want to press this as a counterexample so much as to point out that there is no temptation to generalize Fodor's claim about 'red' to terms less familiar but equally accessible, observationally speaking—terms that should be on a par with 'red' from the point of view of psychophysical law. In exactly the same sense that sufficient conditions for a token's being a 'red'-token can be specified, sufficient conditions for a token's being a 'magenta'-token or a 'Middle C'-token can be specified.

To paraphrase Fodor: there are circumstances such that magenta instantiations control a certain kind of tokening whenever those circumstances obtain; and it's plausible that 'magenta' expresses the property *magenta* in virtue of the fact that magenta instantiations cause tokenings of that kind in those circumstances; and the circumstances are nonsemantically, nonteleologically, and nonintentionally specifiable.(p. 112) Nevertheless, one could be covered by the relevant psychophysical law, see something magenta in the relevant circumstances and still not see it as magenta. I suspect that infants with intact visual systems can see magenta without having the concept MAGENTA.

Similarly, for Middle C. Middle C ought to be susceptible to psychophysical law in the same sense as red is; but to hear a tone *as* Middle C, even if one has "perfect pitch," requires a good deal of auxiliary information beyond what is immediately available in the nonintentionally and nonsemantically described perceptual circumstances. At the least, the disappearance, under optimal circumstances, of a distinction between seeing and seeing-as does not seems to distinguish "psychophysical concepts" from others. Moreover, since there is no reason to think that the concept RED is any less embedded in a system of concepts than is the concept MAGENTA or the concept MIDDLE C, it is not even clear that the distinction between seeing and seeing-as vanishes in the case of RED.

On the other hand, the similarity between 'red' and 'water' is immediately apparent. Molecular duplicates on Earth and on Twin Earth have different concepts: one of WATER and the other of WATER2, by virtue of environmental differences. Molecular duplicates on Earth and on Mars have different concepts: one of RED and the other of GREEN, by virtue of environmental differences (on one reading). The molecular duplicates on Earth and Twin Earth have the same causal powers; the molecular duplicates on Earth and Mars have the same causal powers. 'Red'-tokens are no closer to being "self-interpreting" than are 'water'-tokens. So, there does not seem to be a semantically special class of observation terms anyway; on one side, 'red' does not seem relevantly different from the nonobservation term 'water', and on the other side, what is supposed to make observation terms special clearly does not apply to most so-called observation terms—if it applies to any at all.

If the naturalization problem, under the constraints set out by Fodor, is a problem at all, it thus seems unlikely to be solvable by appeal to a general, context-free distinction between observation terms and other terms. So, if the naturalization problem as it has been recently conceived is genuinely a problem, I believe that we remain at a loss for a solution.[19]

Notes

1. See, for example, Dretske (1981); Stalnaker (1984); Block (1986); and Fodor (1987). Hereafter, references to the latter will be given in the text as 'p.' followed by a page number.
2. Fodor (1984): p. 232.
3. Fodor's usage of a key term is inconsistent. Is "A'-tokening' a semantic characterization that identifies the token in terms of its semantic type (interpretation)? Do 'A'-tokens represent A's? Sometimes yes, sometimes no. On the one hand, when Fodor says that only A's cause (nonwild) 'A'-tokens, the "A'-tokens' is a semantic characterization of tokens as representing A's. And he clearly uses "horse'-tokening' in such a way that 'horse'-tokenings represent horses, and similarly for 'red'-tokenings. On the other hand Fodor sometimes assumes that it is an open questions as to what properties 'A'-tokenings represent; he wants to distinguish the case in which 'A' expresses the property of being A (and hence B-caused 'A'-tokens misrepresent) from the case in which 'A' expresses the property of being A or B (and hence B-caused 'A'-tokens are veridical).(p. 101-102) (I think that it would be less misleading to express the distinction as one between 'A'-tokens and 'A or B'-tokens and to take 'A'-tokens as always representing A's.)

4. Following Fodor, I shall italicize names of properties (the property *red*), put names of concepts in all capitals (the concept RED). "RED is the concept which denotes (or expresses) *red*; and 'red' is a term (either of English or Mentalese) that encodes that concept."(p. 160 5n) On Fodor's view, concepts are expressions of Mentalese. The content of a symbol, as I am using the terms, is what it expresses or denotes or represents.
5. Fodor speaks of the asymmetric dependence of misrepresentation on *truth*; perhaps 'the asymmetric dependence of misrepresentation on accurate representation' would be better.
6. Putnam (1975): 238.
7. Fodor replies to an objection about learning 'horse' from noninstances (cows, say), in a way that requires him to distinguish (as Dretske also must) between training period and mastery period. As Fodor points out in his criticism of Dretske, it is far from clear that there is a principled distinction between the two periods. In any case, the example of the robot case can not be avoided by Fodor's reply to the cited objection. See Fodor's discussion of current dispositions (p. 109).
8. Bernard Kobes pointed out that there is room for asymmetric dependence if we assume that, on the basis of his participation in linguistic and other community practices, *S* learned the concept CAT even though she had only been exposed to robot-cats. Although I think that this is plausible, I do not think that it is available to Fodor. Since on Fodor's view, the meaning of public language is dependent on the meaning of mental representations, invocation of interpreted public language at the ground level of his account of the semantics of mental representation would be viciously circular.
9. P. 106. If, in this context, 'A'-tokenings need not be misrepresentations at all.
10. See, for example, Baker (1987), (1987b).
11. Again see Baker (1987), (1987b).
12. Moreover, referring to Earth's atmosphere in the specification of sufficient conditions would not guarantee that we had truly sufficient conditions anyway. For we could imagine that the inhabitants of all three planets always (even when they are doing their science) wear certain kinds of glasses, but that the glasses have different properties on the different planets. On Earth, they have no effect; on Mars, they transform waves of lengths that green objects on Earth reflect to waves of lengths that red objects on Earth reflect; on Venus, they transform waves of lengths that red objects on Earth reflect to waves of lengths that green objects on Earth reflect. It is implausible to suppose that we could formulate a law that anticipates all the conditions in which same causes produce different effects.
13. Although similar to a kind of converse of Kripke's puzzle about belief, the issues raised here are different. At least as David Lewis sees it, if beliefs are not narrow, Kripke's puzzle vanishes; however, the difficulty raised here does not. Kripke (1979); Lewis (1981): p. 288.
14. It is thus incoherent to say that " 'red' is a term (*either in English or Mentalese*) that encodes [the concept RED]." (p. 160, 5n; my emphasis)

15. For a discussion of related issues, see Pereboom (forthcoming).
16. One source of difficulty is that for (almost?) any specifiable context, there is a larger context that can embed the specified context and alter its effects. So, even if we can specify conditions in which an *X*-instantiation reliably produces an *X*-appearance or an 'X'-token, appeal to the counterfactual-supporting properties of causal statements licenses consideration of a larger set of conditions (which encompass the original set) in which it is implausible or even incoherent to claim that an *X*-instantiation produces an *X*-appearance. Although my argument calls into question all forms of mind/brain supervenience, as long as minds are characterized by intentional states like belief, it does not apply to the kind of "weak supervenience" that John Haugeland (1982) has advocated.
17. Bernard Kobes suggested re-description of the cases along the lines of (c).
18. Since I think that variants of (A1) and (A2) both have great intuitive appeal, I think that the conflict presented here is evidence of a deep problem, one that I do not yet know how to resolve. My point here is only that the Causal Theory of Content, which tends to pull away from mind/brain supervenience, is no solution.
19. I read versions of this paper at the University of Massachusetts at Amherst and at Arizona State University, where I received incisive comments, especially from Gary Hardegree, Gareth B. Matthews, Bernard Kobes, Gregory Fitch and Theodore Guleserian. I also wish to thank Hilary Kornblith for helping to improve an earlier draft and to acknowledge the Middlebury College Faculty Research Fund for support of this work.

References

Baker, Lynne Rudder (1987), "Content by Courtesy," *Journal of Philosophy* 84: 197-213.

Baker, Lynne Rudder (1988), *Saving Belief* (Princeton: Princeton University Press).

Block, Ned (1986), "Advertisement for a Semantics for Psychology," in *Studies in the Philosophy of Mind* (Midwest Studies in Philosophy, Vol. X), Peter A. French, Theodore E. Uehling, Jr., Howard K. Wettstein, eds. (Minneapolis: University of Minnesota): 615-678.

Dretske, Fred I. (1981), *Knowledge and the Flow of Information* (Cambridge, Mass: MIT/Bradford).

Fodor, Jerry A. (1984), "Semantics Wisconsin Style," *Synthese* 59: 231-250.

Fodor, Jerry A. (1987), *Psychosemantics* (Cambridge, Mass: MIT/Bradford).

Haugeland, John (1982), "Weak Supervenience," *American Philosophical Quarterly* 19: 93-103.

Kripke, Saul A. (1979), "A Puzzle About Belief," in *Meaning and Use*, Avishai Margalit, ed. (Dordrecht/Holland: D. Reidel Publishing Co.): 239-283.

Lewis, David (1981), "What Puzzling Pierre Does Not Believe," *Australasian Journal of Philosophy* 59 (1981): 283-289.

Pereboom, Derk (forthcoming), "Why a Realist Cannot Be a Functionalist."

Putnam, Hilary (1975), "It Ain't Necessarily So" in *Mathematics, Matter and Method: Philosophical Papers, Vol. I* (Cambridge, Cambridge University): 237-249

Stalnaker, Robert C. (1984), *Inquiry* (Cambridge, Mass: MIT/Bradford).