J. A. FODOR

SEMANTICS, WISCONSIN STYLE

There are, of course, two kinds of philosophers. One kind of philosopher takes it as a working hypothesis that belief/desire psychology (or, anyhow, *some* variety of propositional attitude psychology) is the best theory of the cognitive mind that we can now envision; hence that the appropriate direction for psychological research is the construction of a belief/desire theory that is empirically supported and methodologically sound. The other kind of philosopher takes it that the entire apparatus of propositional attitude psychology is conceptually flawed in irremediable ways; hence that the appropriate direction for psychological research is the construction of alternatives to the framework of belief/desire explanation. This way of collecting philosophers into philosopher-kinds cuts across a number of more traditional, but relatively superficial, typologies. For example, eliminativist behaviorists like Quine and neurophiles like the Churchlands turn up in the same basket as philosophers like Steve Stich, who think that psychological states are computational and functional all right, but not intentional. Dennett is probably in that basket too, along with Putnam and other (how should one put it?) dogmatic relativists. Whereas, among philosophers of the other kind one finds a motley that includes, very much *inter alia*, reductionist behaviorists like Ryle and (from time to time) Skinner, radical individualists like Searle and Fodor, mildly radical anti-individualists like Burge, and, of course, all cognitive psychologists except Gibsonians.

Philosophers of the first kind disagree with philosophers of the second kind about many things besides the main issue. For example, they tend to disagree vehemently about who has the burden of argument. However – an encouraging sign – recent discussion has increasingly focused upon one issue as the crux par excellence on which the resolution of the dispute must turn. The point about propositional attitudes is that they are *representational* states: Whatever else a belief is, it is a kind of thing of which semantic evaluation is appropriate. Indeed, the very individuation of beliefs proceeds via (oblique) reference to the states of affairs that determine their semantic value; the

Synthese **59** (1984) 231–250. 0039–7857/84/0593–0231 $02.00

belief that it is raining is essentially the belief whose truth or falsity depends on whether it is raining. Willy nilly, then, the friends of propositional attitudes include only philosophers who think that serious sense can be made of the notion of representation (de facto, they tend to include *all* and only philosophers who think this). I emphasize that the notion of representation is crucial for every friend of propositional attitudes, not just the ones (like, say, Field, Harman and Fodor) whose views commit them to quantification over symbols in a mental language. Realists about propositional attitudes are *ipso facto* Realists about representational states. They must therefore have *some* view about what it is for a state to *be* representational even if (like, say, Loar and Stalnaker) they are agnostic about, or hostile towards, identifying beliefs and desires with sentences in the language of thought.

Well, what would it be like to have a serious theory of representation? Here, too, there is some consensus to work from. 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. What is required to relieve the worry is therefore, at a minimum, the framing of *naturalistic* conditions for representation. That is, what we want at a minimum is something of the form '*R represents S*' *is true iff C* where the vocabulary in which condition C is couched contains neither intentional nor semantical expressions.[1,2]

I haven't said anything, so far, about what R and S are supposed to range over. I propose to say as little about this as I can get away with, both because the issues are hard and disputatious and because it doesn't, for the purposes of this paper, matter much how they are resolved. First, then, I propose to leave it open which things *are* representations and how many of the things that qualify a naturalistic theory should cover. I assume only that we must have a naturalistic treatment of the representational properties of the propositional attitudes; if propositional attitudes are relations to mental representations, then we must have a naturalistic treatment of the representational properties of the latter.

In like spirit, I propose to leave open the ontological issues about the possible values of S. The paradigmatic representation relation I have in mind holds between things of the sorts that have truth values and things of the sorts by which truth values are determined. I shall usually refer to

the latter as "states of affairs", and I'll use '-ing nominals' as canonical forms for expressing them (eg., 'John's going to the store'; 'Mary's kissing Bill'; 'Sam's being twelve years old next Tuesday'). Since the theories we'll discuss hold that the relations between a representation and what it represents are typically causal, I shall assume further that S ranges over kinds of things that can *be* causes.

Last in this list of things that I'm not going to worry about is type token ambiguities. A paradigm of the relation we're trying to provide a theory for is the one that holds between my present, occurrent belief that Reagan is president and the state of affairs consisting of Reagan's being President. I assume that this is a relation between tokens; between an individual belief and an individual state of affairs. But I shall also allow talk of relations between representation *types* and state of affair *types*; the most important such relation is the one that holds when tokens of a situation type cause, or typically cause, tokenings of a representation type. Here again there are ontological deep waters; but I don't propose to stir them up unless I have to.

OK, let's go. There are, so far as I know, only two sorts of naturalistic theories of the representation relation that have ever been proposed. And at least one of these is certainly wrong. The two theories are as follows: that *C* specifies some sort of *resemblance* relation between *R* and *S*: and that *C* specifies some sort of *causal* relation between *R* and *S*.[3] The one of this pair that is certainly wrong is the resemblance theory. For one thing, as everybody points out, resemblance is a symmetrical relation and representation isn't; so resemblance can't *be* representation. And, for another, resemblance theories have troubles with the *singularity* of representation. The concept *tiger* represents *all tigers*; but the concept *this tiger* represents only this one. There must be (possible) tigers that resemble this tiger to any extent you like, and if resemblance is sufficient for representation, you'd think the concept *this tiger* should represent those tigers too. But it doesn't, so again resemblance can't be sufficient for representation.

All this is old news. I mention it only to indicate some of the ways in which the idea of a causal theory of representation is prima facie attractive, and succeeds where resemblance theories fail. (1) Causal relations are natural relations if *anything* is. You might wonder whether resemblance is part of the natural order (or whether it's only, as it were, in the eye of the beholder). But to wonder that about causation is to wonder whether there *is* a natural order. (2) Causation, unlike resem-

blance, is nonsymmetric, (3) Causation is par exellence, a relation among *particulars*. Tiger *a* can resemble tiger *b* as much as you like, and it can still be tiger *a* and not tiger *b* that caused this set of tiger prints. Indeed, if it was tiger *a* that caused them, it *follows* that tiger *b* didn't (assuming, of course, that tiger *a* is distinct from tiger *b*).

Well, in light of all this, several philosophers who are sympathetic towards propositional attitudes have recently been playing with the idea of a causal account of representation (see, particularly, Stampe (1975; 1977), Dretske (1981; forthcoming) and Fodor (forthcoming). Much of this has been going on at the University of Wisconsin, hence the title of this essay.) My present purpose is to explore some consequences of this idea. Roughly, here's how the argument will go: causal theories have trouble distinguishing the conditions for *representation* from the conditions for *truth*. This trouble is intrinsic; the conditions that causal theories impose on representation are such that, when they're satisfied, *mis*representation cannot, by that very fact, occur. Hence, causal theories about how propositional attitudes represent have Plato's problem to face: how is false belief possible? I'll suggest that the answer turns out to be that, in a certain sense, it's not, and that this conclusion may be more acceptable than at first appears.

I said I would argue for all of that; in fact I'm going to do less. I propose to look at the way the problem of misrepresentation is handled in the causal theories that Stampe and Dretske have advanced; and I really *will* argue that their treatments of misrepresentation don't work. This exercise should make it reasonably clear why misrepresentation is so hard to handle in causal theories generally. I'll then close with some discussion of what we'll have to swallow if we choose to bite the bullet. The point of all this, I emphasize, is *not* to argue against causal accounts of representation. I think, in fact, that something along the causal line is the best hope we have for saving intentionalist theorizing, both in psychology and in semantics. But I think too that causal theories have some pretty kinky consequences, and it's these that I want to make explicit.

To start with, there are, strictly speaking, *two* Wisconsin theories about representation; one that's causal and one that's epistemic. I propose to give the second pretty short shrift, but we'd better have a paragraph or two.

The basic idea of (what I shall call) an epistemic access theory is that R represents S if you can find out about S from R.[4] So, for example,

Dretske says (EB, 10) "A message . . . carries information about X to the extent to which one could learn (come to know) something about X from the message." And Stampe says (S&T 223): "An object will represent or misrepresent the situation . . . only if it is such as to enable one to come to know the situation, i.e., what the situation is, should it be a faithful representation."

Now, generally speaking, if representation requires that S cause R, then it will of course be possible to learn about R by learning about S; inferring from their effects is a standard way of coming to know about causes. So, depending on the details, it's likely that an epistemic account of representation will be satisfied whenever a causal one is. But there is no reason to suppose that the reverse inference holds, and we're about to see that epistemic accounts have problems to which the causal ones are immune.

(1) The epistemic access story (like the resemblance story) has trouble with the nonsymmetry of representation. You can find out about the weather from the barometer, but you can also find out about the barometer from the weather since, if it's storming, the barometer is likely to be low. Surely the weather doesn't represent the barometer, so epistemic access can't be sufficient for representation.

(2) The epistemic story (again like the one about resemblance) has trouble with the singularity of representation. What shows this is a kind of case that Stampe discusses extensively in TCTLR. Imagine a portrait of, say, Chairman Mao. If the portrait is faithful, then we can infer from properties of the picture to properties of the Chairman (e.g., if the portrait is faithful, then if it shows Mao as bald, then we can learn *from* the portrait *that* Mao is bald). The trouble is, however, that if Mao has a Doppelgänger and we know he does, then we can *also* learn from the portrait that Mao's Doppelgänger is bald. But the portrait is of Mao and not of his Doppelgänger for all that.

Dretske has a restriction on his version of the epistemic access theory that is, I expect, intended to cope with the singularity problem; he allows that a message carries information about X only if a "suitably equipped but *otherwise ignorant* receiver" could learn about X from the message (EB 10, my emphasis). I imagine the idea is that, though we could learn about Mao's Doppelgänger from Mao's portrait, we couldn't do so *just from the portrait alone*; we'd also have to use our knowledge that Mao has a Doppelgänger. I doubt, however, that this further condition can really be enforced. What Dretske has to face is, in effect,

the Dreaded Collateral Information Problem; i.e., the problem of how to decide when the knowledge that we use to interpret a symbol counts as knowledge about the symbol, and when it counts as collateral knowledge. This problem may seem self-solving in the case of *pictures* since we have a pretty good pretheoretical notion of which properties of a picture count as the pictorial ones. But in the case of, e.g., linguistic symbols, it's very far from evident how, or even whether, the corresponding distinction can be drawn. If I say to you "John is thirty two", you can learn something reliable about John's age from what I said. But, of course, you can also learn something reliable about John's *weight* (e.g., that he weighs more than a gram). It may be possible to discipline the intuition that what you learn about John's age you learn just from the symbol and what you learn about his weight you learn from the symbol plus background information. But drawing that distinction is notoriously hard and, if the construal of representation depends on our doing so, we are in serious trouble.

(3) Epistemic theories have their own sorts of problems about misrepresentation. Stampe says,

> An object will represent or misrepresent the situation . . . only if it is such as to enable one to come to know the situation, i.e., what the situation is, should it be a faithful representation. If it is not faithful, it will misrepresent the situation. That is, one *may* not be able to tell from it what the situation is, despite the fact that it is a representation of the situation. In either case, it represents the same thing, just as a faithful and an unrecognizable portrait may both portray the same person.

But, to begin with, the example is perhaps a little question-begging, since it's not clear that the bad portrait represents its sitter *in virtue of* the fact that if it were accurate it would be possible to learn from it how the sitter looks. How, one wonders, could this bare counterfactual determine representation? Isn't it, rather, the other way around; i.e., not that it's a portrait of Mao because (if it's faithful) you can find out about Mao from it, but rather that you can find out about Mao from it (if it's faithful) because it's Mao that it's a portrait of.

To put the same point slightly differently: we'll see that causal theories have trouble saying how a symbol could be tokened and still be false. The corresponding problem with epistemic access theories is that they make it hard to see how a symbol could be *intelligible* and false. Stampe says: "An object will represent or misrepresent the situation . . . only if it is such as to enable one to come to know the situation, i.e., what the situation is, should it be a faithful represen-

tation." (S&T 223). Now, there is a nasty scope ambiguity in this; viz, between:

(a) if R is faithful (you can tell what the case is); vs.

(b) you can tell (what the case is if R is faithful).

It's clear that it is (a) that Stampe intends; ((b) leads in the direction of a possible world semantics, which is where Stampe explicitly doesn't want to go; see especially SAT, circa p. 224). So, consider the symbol "Tom is Armenian", and let's suppose the fact – viz., the fact in virtue of which that symbol has its truth value – is that Tom is Swiss. Then Stampe wants it to be that what the symbol represents (i.e., *mis*-represents) is Tom's being Swiss; *that's* the fact to which, if it were faithful, the symbol would provide epistemic access.

Now, to begin with, this counterfactual seems a little queer. What, precisely, would it be *like* for "Tom is Armenian" to be faithful to the fact it (mis)represents – viz., to the fact that Tom is Swiss? Roughly speaking, you can make a false sentence faithful either by changing the world or by changing the sentence; but neither will do the job that Stampe apparently wants done.

(1) Change the world: make it be that Tom is Armenian. The sentence is now faithful, but to the wrong fact. That is, the fact that it's now faithful to isn't the one that it (mis)represented back when it used to be untrue; that, remember, was the fact that Tom is Swiss.

(2) Change the sentence: make it *mean* that Tom is Swiss. The sentence is now faithful to the fact that it used to (mis)represent. But is the counterfactual intelligible? Can we make sense of talk about what a sentence would represent if it – the very same sentence – meant something different? And, if meaning can change while what is represented stays the same, in what sense does a theory of representation constitute a theory of meaning?

Problems, problems. Anyhow, the main upshot is clear enough, and it's one that Stampe accepts. According to the epistemic access story, when a symbol *mis*represents, "one *may* not be able to tell from it what the situation is, despite the fact that it is a representation of the situation". Here not being "able to tell what the situation is" doesn't mean not being able to tell what it is that's *true* in the situation; it means not being able to tell *what situation it is that the symbol represents*. You can't tell, for example, that the symbol "Tom is Armenian" represents

Tom's being French unless you happen to know Tom's nationality.

It may be supposed that Stampe could disapprove of this along the following lines: you *can*, in one sense, tell what "Tom is Armenian" represents even if you don't know that Tom is Swiss. For, you can know that "Tom is Armenian" represents Tom's nationality (i.e., that if it's faithful it provides epistemic access to his nationality) even if you don't know what Tom's nationality is. I think this is OK, but you buy it at a price: On this account, knowing what a symbol *represents* (what it provides epistemic access to) can't be equated with knowing what the symbol *means*. Notice that though "Tom is Armenian" has the property that if it's faithful it provides epistemic access to Tom's nationality, so too do a scillion other, nonsynonymous sentences like "Tom is Dutch", "Tom is Norwegian", "Tom is Swiss", and so forth. To put the same point another way, on the present construal of Stampe's account, what a truth valuable symbol represents isn't, in general, its truth condition. (The truth condition of a symbol is the state of affairs which, if it obtains, would make the symbol true; and what would make "Tom is Armenian" true is Tom's being Armenian, not Tom's being Swiss.) Correspondingly, what you can know about "Tom is Armenian" if you don't know that Tom is Swiss is not what its truth condition is, but only what it represents; viz., that it represents Tom's nationality. This means that Stampe has either to give up on the idea that understanding a symbol is knowing what would make it true, or develop a reconstruction of the notion of truth condition as well as a reconstruction of the notion of representation. Neither of these alternatives seems particularly happy.

There's more to be said about the epistemic approach to representation; but let's, for present purposes, put it to one side. From here on, only causal accounts will be at issue.

The basic problem for causal accounts is easy enough to see. Suppose that S is the truth condition of R in virtue of its being the cause of R. Now, causation is different from resemblance in the following way; a symbol can (I suppose) resemble something merely possible; it's OK for a picture to be a picture of a unicorn. But, surely, no symbol can be an effect of something merely possible. If S causes R, then S obtains. But if S obtains and S is the truth condition of R, it looks as though R has to be true; being true just *is* having truth conditions that obtain. So it looks like this: a theory that numbers *causation* among the relations in virtue of which a representation has its truth conditions is going to allow

truth conditions to be assigned only when they're satisfied. I don't say that this argument is decisive; but I do say – and will now proceed to argue – that Wisconsin semantics hasn't thus far found a way around it.

I'll start with Dretske's treatment of the misrepresentation problem in *Knowledge And the Flow of Information*. The crucial passage is on pp. 194–195. Here is what Dretske says:

> In the learning situation special care is taken to see that incoming signals have an intensity, a strength, sufficient unto delivering the required piece of information *to* the learning subject... Such precautions are taken in the learning situation... in order to ensure that an internal structure is developed with the information that *s* is *F*... But once we have meaning, once the subject has articulated a structure that is selectively sensitive to information about the *F*-ness of things, instances of this structure, tokens of this type, can be triggered by signals that *lack* the appropriate piece of information... We (thus) have a case of misrepresentation – a token of a structure with a false content. We have, in a word, meaning with truth. (Emphasis Dretske's.)

All you need to remember to understand this well enough for present purposes is (1) that Dretske's notion of information is fundamentally that of counterfactual supporting correlation (i.e., that objects of type *R* carry information about states of affairs of type *S* to the extent that tokenings of the type *S* are nomically responsible for tokenings of the type *R*). And (2) that the tokening of a representation carries the information that *s* is *F* in *digital* form if and only if the information that *s* is *F* is the most specific information that that tokening carries about *s*. Roughly speaking, the pretheoretic notion of the *content* of a representation is reconstructed as the information that the representation digitalizes.

Now then: how does *mis*representation get into the picture? There is, of course, no such thing as mis*information* on Dretske's sort of story. Information is correlation and though correlations can be better or worse – more or less reliable – there is no sense to the notion of a *mis*correlation: hence there is nothing, so far, to build the notion of misrepresentation out of.

The obvious suggestion would be this: suppose *R*s are nomically correlated with – hence carry information about – *S*s; then, as we've seen, given the satisfaction of further (digitalization) conditions, we can treat *R*s as representations of *S*s: S is the state of affairs type that symbols of the *R* type represent. But suppose that, from time to time, tokenings of *R* are brought about (not by tokenings of *S* but) in some *other* way. Then these, as one might say, 'wild' tokenings would count

as *mis*representations: for, on the one hand, they have the content that S; but, on the other hand, since it isn't the fact that S that brings about their tokening the content that they have is false. *Some* sort of identification of misrepresentations with etiologically wild tokenings is at the heart of all causal accounts of misrepresentation.

However, the crude treatment just sketched clearly won't do: it is open to an objection that can be put like this: If there are wild tokenings of R, it follows that the nomic dependence of R upon S is imperfect; some R-tokens – the wild ones – are *not* caused by S tokens. Well, but clearly they are caused by *something*; i.e., by something that is, like S, sufficient but not necessary for bringing Rs about. Call this second sort of sufficient condition the tokening of situations of type T. Here's the problem: R represents the state of affairs with which its tokens are causally correlated. Some representations of type R are causally correlated with states of affairs of type S; some representations of type R are causally correlated with states of affairs of type T. So it looks as though what R represents is not either S or T, but rather the disjunction $(S \vee T)$: The correlation of R with the disjunction is, after all, *better than* its correlation with either of the disjuncts and, ex hypothesi, correlation makes information and information makes representation. If, however, what Rs represent is not S but $(S \vee T)$, then tokenings of R that are caused by T *aren't, after all, wild tokenings* and our account of misrepresentation has gone West.

It is noteworthy that this sort of argument – which, in one form or other, will be with us throughout the remainder of this essay – seems to be one that Dretske himself accepts. The key assumption is that, ceteris paribus, if the correlation of a symbol with a disjunction is better than its correlation with either disjunct, it is the disjunction, rather than either disjunct, that the symbol represents. This is a sort of "principle of charity" built into causal theories of representation: 'so construe the content of a symbol that what it is taken to represent is what it correlates with *best*'. Dretske apparently subscribes to this. For example, in EB (circa p. 17) he argues that, for someone on whose planet there is both XYZ and H_2O but who learns the concept *water* solely from samples of the former, the belief that such and such is water is the belief that it is that it is *either* H_2O *or* XYZ. This seems to be charity in a rather strong form: R represents a disjunction even if all tokenings of R are caused by the satisfaction of the *same* disjunct, so long as satisfaction of the other disjunct *would have caused R tokenings had*

they happened to occur. I stress this by way of showing how much the counterfactuals count; Dretske's conditions on representation are intensional (with an '*s*'); they constrain the effects of counterfactual causes.

To return to Dretske's treatment of misrepresentation: his way out of the problem about disjunction is to enforce a strict distinction between what happens in the learning period and what happens after. Roughly, the correlations that the learning period establish determine what *R* represents; and the function of the Teacher is precisely to insure that the correlation so established is a correlation of *R* tokens with *S* tokens. It may be that *after* the learning period, *R* tokens are brought about by something *other than S* tokens; if so, these are wild tokenings of *R* and their contents are false.

This move is ingenious but hopeless. Just for starters, the distinction between what happens in the learning period and what happens thereafter surely isn't principled; there is no time after which one's use of a symbol stops being merely shaped and starts to be, as it were, in earnest. Perhaps idealization will bear some of this burden, but it's hard to believe that it could yield a notion of learning period sufficiently rigorous to underwrite the distinction between truth and falsity; which is, after all, precisely what's at issue. Second, if Dretske does insist upon the learning period gambit, he limits the applicability of his notion of misrepresentation to *learned* symbols. This is bad for me because it leaves us with no way in which innate information could be false; and it's bad for him because it implies a basic dichotomy between *natural* representation (smoke and fire; rings in the tree and the age of the tree) and the intentionality of mental states.

All of that, however, is mere limbering up. The real problem about Dretske's gambit is internal; it just doesn't work. Consider a trainee who comes to produce *R* tokens in *S* circumstances during the training period. Suppose, for simplification, that the correlation thus engendered is certainly nomic, and that *S* tokenings are elicited by *all and only R* tokenings during training: error-free learning. Well, time passes, a whistle blows (or whatever), and the training period comes to an end. At some time later still, the erstwhile trainee encounters a tokening of a *T* situation (*T* not equal to *S*) and produces an *R* in causal consequence. The idea is, of course, that this *T*-elicited tokening of *R* is ipso facto wild and, since it happens after the training period ended, it has the (false) content *that S*.

But, as I say, this won't work: it ignores relevant counterfactuals. Imagine, in particular, what *would have* happened if a token of situation type *T had* occurred during the training period. Presumably what would have happened is that it would have elicted a tokening of *R*. After all, tokenings of *T* are assumed to be sufficient to cause *R* tokenings *after* training; that's the very assumption upon which Dretske's treatment of wild *R*-tokenings rests. So we can assume – indeed, we can stipulate – that *T* is a situation which, if it had occured *during* training, would have been sufficient for *R*. But that means, of course, that if you include the counterfactuals, the correlation that training established is (not between *R* and *S* but) between *R* and the disjunction $(S \vee T)$. So now we have the old problem back again. If training established a correlation with $(S \vee T)$ then the content of a tokening of *R* is *that* $(S \vee T)$. So a tokening of *R* caused by *T* isn't a wild tokening after all; and since it isn't wild it also isn't false. A token with the content $(S \vee T)$ is, of course, *true* when it's the case that *T*.

There is a sort of way out for Dretske. He could say this: 'The trouble is, you still haven't taken care of *all* the relevant counterfactuals; in particular, you've ignored the fact that if a *T*-tokening has occurred during training and elicited an *R*-tokening *the Teacher would have corrected the R response*. This distinguishes the counterfactual consequences of *T*-elicited *R*-tokens occuring during training from those of *S*-elicited *R*-tokens occuring during training since the latter would not, of course, have been corrected. In the long run, then, it is *these* counterfactuals – ones about what the teacher *would have corrected* – that are crucial; *R*s represent *S*s (and not *T*s) because the Teacher would have disapproved of *T*-elicited *R*-responses if they had occurred.'

But I don't think Dretske would settle for this, and nor will I. It's no good for Dretske because it radically alters the fundamental principle of his theory, which is that the character of symbol-to-situation correlations determines the content of a symbol. On this revised view, the essential determinant is not the actual, or even the counterfactual, correlations that hold between the symbol and the world; rather it's the Teacher's pedagogical intentions; specifically, the Teacher's intention to reward only such *R* tokenings as are brought about by *S*s. And it's no good for me because it fails a prime condition upon *naturalistic* treatment of representations; viz, that appeals to intentional (with a '*t*') states must not figure essentially therein. I shall therefore put this

suggestion of Dretske's to one side and see what else may be on offer.

Let's regroup. The basic problem is that we want there to be conditions for the *truth* of a symbol over and above the conditions whose satisfaction determines what the symbol represents. Now, according to causal theories, the latter – representation determining – conditions include whatever is necessary and suffcient to bring about tokenings of the symbol (including nomically possible counterfactual tokenings.) So the problem is, to put it crudely, if we've already used up all that to establish representation, what more could be required to establish truth?

An idea that circulates in all the texts I've been discussing (including my own) goes like this. Instead of thinking of the representation making conditions as whatever is necessary and sufficient for causing tokenings of the symbol, think of them as whatever is necessary and sufficient for causing such tokenings *in normal circumstances*. We can then think of the wild tokens as being (or, anyhow, as including) the ones which come about when the 'normal conditions' clause is *not satisfied*. This doesn't, of course, get us out of the woods. At a minimum, we still need to show (what is by no means obvious) that for a theory of representations to appeal to normalcy conditions (over and above causal ones) isn't merely question-begging; for example, that you can characterize what it is for the conditions of a tokening to be normal without invoking intentional and/or semantic notions. Moreover, we'll also have to show that appealing to normalcy conditions is a way of solving the disjunction problem; and that, alas, isn't clear either. We commence with the first of these worries.

It is, I think, no accident that there is a tendency in all the texts I've been discussing (again including mine) to introduce normalcy conditions by appeal to examples where *teleology* is in play. For example, to use a case that Dretske works hard, a voltmeter is a device which, under normal conditions, produces an output which covaries (nomically) with the voltage across its input terminals. 'Normal conditions' include that all sorts of constraints on the internal and external environment of the device should be satisfied (e.g., the terminals must not be corroded) but it seems intuitively clear that what the device registers is the voltage and not the voltage together with the satisfaction of the normalcy conditions. If the device reads zero, that means that there's no current flowing, not that *either* there is no voltage flowing *or* the terminals are

corroded.

However, we know this because we know what the device is *for* and we can know what the device is for only because there *is* something that the device is for. The tendency of causal theorists to appeal to teleology for their best cases of the distinction between representation-making causal conditions and mere normalcy conditions is thus unnerving. After all, in the case of artifacts at least, being 'for' something is surely a matter of being *intended* for something. And we had rather hoped to detach the representational from the intentional since, if we can't, our theory of representation ipso facto fails to be naturalistic and the point of the undertaking becomes, to put it mildly, obscure.

There are, it seems, two possibilities. One can either argue that there can be normalcy without teleology (i.e., that there are cases *other than* teleological ones where a distinction between causal conditions and normal conditions can be convincingly drawn); or one can argue that there can be teleology without intentionality (*natural* teleology, as it were) and that the crucial cases of representation rest exclusively upon teleology of this latter kind. Unlike Dretske and Stampe, I am inclined towards the second strategy. It seems to me that our intuitions about the distinction between causal and normal conditions are secure only in the cases where the corresponding intuitions about teleology are secure, and that wherever we *don't* have intuitions about teleology, the disjunction argument seems persuasive.[5] Let's look at a couple of cases.

Thermometers are OK; given normalcy conditions (e.g., a vacuum in the tube) the nomic covariance between the length of the column and the temperature of the ambient air determines what the device represents. Violate the normalcy conditions and, intuition reports, you get wild readings; i.e., *mis*representations of the temperature. But, of course, thermometers are *for* measuring something, and precisely what they're for measuring (viz., the temperature of the ambient air) is what the present analysis treats as a causal (rather than a normalcy) condition. Compare, by way of contrast, the diameter of the coin in my pocket. Fix my body temperature and it covaries with the temperature of the ambient air; fix the temperature of the ambient air, and it covaries with the temperature of my body. I see *no* grounds for saying that one of these things is what it really represents and the other is a normalcy condition (e.g., that the diameters that are affected by body temperature are misrepresentations of the air temperature).[6] In short, where there is no question of teleology it looks as though one's

intuitions about which are the normalcy conditions are unstable. Such examples should make one dubious about the chances for a notion of normalcy that applies in *non*teleological cases.

Or, consider an example of Stampe's: (CTCLR, 49)

> The number of rings in (a tree stump) represents the age of the tree... The causal conditions, determining the production of this representation, are most saliently the climatic conditions that prevailed during the growth of the tree. If these are normal... then one ring will be added each year. Now what *is* that reading... It is not, for one thing, infallible. There may have been drought years... It is a *conditional* hypothesis: that *if* certain conditions hold, then something's having such and such properties would cause the representation to have such and such properties... Even under those normal conditions, there may be other things that would produce the rings – an army of some kind of borer, maybe, or an omnipotent evil tree demon.

Stampe's analysis of this case rests on his decision to treat the seasonal climatic variations as the causal component of the conditions on representation and the absence of (e.g.) drought, tree borers, evil demons and the rest as normalcy conditions. And, of course, given that decision, it's going to follow from the theory that the tree's rings represent the tree's age and that tree-borer-caused tree ring tokens are wild (i.e., that they *mis*represent the tree's age). The worrying question is what, if anything, motivates this decision.

We should do this in several steps. Let's consider a particular case of tree-borer-caused tree ring tokenings. Suppose, for the moment, we agree that the general truth is that a tree's rings represent the tree's age. And suppose we agree that it follows from this general truth that all tree ring tokenings represent the age of the tree that they're tokened in. Well, even given all that it's not obvious what these tree-borer-caused tokenings represent since it's not obvious that they are, in the relevant sense, tree rings.

Perhaps the right way to describe the situation is to say that these things merely *look like* tree rings. Compare the token of "Look upon my works, oh ye mighty, and despair" that the wind traces in the desert sands. This *looks like* a token of an English sentence type (and, of course, if it *were* a token of that sentence type it would be unfaithful, what with there not being anything to look at and all). But it's not a token of that English sentence since it's not a token of *any* sentence. A fortiori, it's not a wild or unfaithful token. Similarly, mutatis mutandis (maybe) with the putative tree rings; they're not wild (unfaithful) representations of the tree's age because, even if all tree rings are

representation of a tree's age, *these aren't tree rings*.

I hope I will be seen not to be merely quibbling. Stampe wants it to come out that tree-borer caused tree rings are wild; that they're misrepresentations of the tree's age. He needs this a lot since this sort of case is Stampe's paradigm example of a distinction between causal conditions and normalcy conditions which doesn't rest on teleology. But I claim that the case doesn't work *even assuming what's yet to be shown, viz., that tree rings represent tree age rather than tree-age-plus-satisfaction-of-normalcy-conditions*. For Stampe is assuming a nonquestion begging – hence naturalistic – criterion for something being a token of a representation type. And there isn't one. (Of course, we do have a criterion which excludes the wind token's being a sentence inscription; but that criterion is *non*naturalistic, hence unavailable to a causal theorist; it invokes the intentions of the agent who produced the token.)

Now let's look at it the other way. Suppose that these tree-borer caused rings *are* tree rings (by stipulation) and let's ask what they represent. The point here is that even if "under normal conditions, tree rings represent the tree's age" is true, it *still* doesn't follow that *these ab*normally formed tree rings represent the tree's age. Specifically, it doesn't follow that these rings represent the tree's age rather than the tree borer's depridations. (Look closely and you'll see the marks their little teeth left. Do those represent the tree's age too?) This is just the disjunction problem over again, though it shows an interesting wrinkle that you get when you complicate things by adding in normalcy conditions. "If circumstances are normal, *x*s are *F*" doesn't, of course, tell you about the *F*ness of *x*s when circumstances are *ab*normal. The most you get is a counterfactual, viz., "if circumstances *had been* normal, this *x* would have been *F*." Well, in the present case, if etiological circumstances had been normal, these rings would have represented the tree's age (viz., accurately). It doesn't follow that, given the way the etiological circumstances actually were, these rings still represent the tree's age (viz., *in*accurately). What you need is some reason to suppose that etiologically abnormal (hence wild) rings represent the same thing that etiologically normal rings do. This is precisely equivalent to saying that what you need is a solution to the disjunction problem, and that is precisely what I've been arguing all along that we haven't got.

We *would* have it, at least arguably, if this were a teleological case.

Suppose that there is some mechanism which (not only produces tree rings but) produces tree rings with an end in view. (Tree rings are, let's suppose, Mother Nature's calendar). Then there is a trichotomous distinction between (a) tree rings produced under normal circumstances; (b) wild tree rings (inscribed, for example, when Mother Nature is a little tipsy); and (c) things that look like tree rings but aren't (tree borer's depridations). This *does* enforce a distinction between representation, misrepresentation and nonrepresentation; not so much because it relativizes representation to *normalcy*, however, but because it relativizes representation to *end-in-view*. The reason that wild tree rings represent the same things as normal ones is that *the wild ones and the normal ones are supposed to serve the same function*. Notice that it's the intensionality of "supposed to" that's doing all the work.

I'm afraid what all this comes to is that the distinction between normal and wild tokens rests – so far at least – on a pretty strong notion of teleology. It's only in the teleological cases that we have any way of justifying the claim that wild tokens represent the same thing that etiologically normal ones do; and it is, as we've seen, that claim on which the present story about misrepresentation rests. How bad is this? Well, for one thing, it's not as bad as if the distinction had turned out to rest on an *intentional* notion. There are, as I remarked above, plausible cases of nonintentional, natural teleology and a naturalistic theory of representation can legitimately appeal to these. On the other hand, if the line of argument we have been exploring is right, then the hope for a *general* theory of representation (one that includes tree rings, for example) is going to have to be abandoned. Tree rings will have to represent only at a remove, via the interests of an observor, since only what has natural teleology can represent absolutely. This is, as a matter of fact, OK with me. For I hold that only sentences in the language of thought represent in, as it were, the first instance; and they represent in virtue of the natural teleology of the cognitive mechanisms. Propositional attitudes represent *qua* relations to sentences in the language of thought. All other representation depends upon the propositional attitudes of symbol users.

Even allowing all this, however, it is arguable that we haven't yet got a notion of misrepresentation robust enough to live with. For we still have this connection between the etiology of representations and their truth values: representations generated in teleologically normal circumstances must be true. Specifically, suppose M is a mechanism the

function of which is to generate tokens of representation type *R* in, and only in, tokens of situation type *S*; *M* mediates the causal relation between *S*s and *R*s. Then we can say that *M*-produced tokens of *R* are wild when *M* is functioning abnormally; but when *M* is functioning normally (i.e., when it's tokening of *R* is causally contingent, in the right way, upon the tokening of *S*) then not only do the tokens of *R* have the content *that S*, but also the contents of these tokens are satisfied, and what the tokens say is true.

Well, consider the application to belief fixation. It looks as though (1) only beliefs with abnormal etiologies can be false, and (2) 'abnormal etiology' will have to be defined with respect to the teleology of the belief-fixing (i.e., cognitive) mechanisms. As far as I can see, this is tantamount to: "beliefs acquired under epistemically optimal circumstances must be true" since, surely, the function of the cognitive mechanisms will itself have to be characterized by reference to the beliefs it *would* cause one to acquire *in* such optimal circumstances. (I take it for granted that we can't, for example, characterize the function of the cognitive mechanisms as the fixation of *true* beliefs because truth is a semantical notion. If our theory of representation is to rest upon the teleology of the cognitive mechanisms, cognitive teleology must itself be describable naturalistically; viz., without recourse to semantic concepts. For an extended discussion of this sort of stuff, see my op cit.)

It appears that we have come all this way only in order to rediscover verificationism. For, I take it, verificationism just *is* the doctrine that truth is what we would believe in cognitively optimal circumstances. Is this simply too shameful for words? Can we bear it? I have three very brief remarks to make. They are, you will be pleased to hear, concluding remarks.

First, *all* Naturalistic theories in semantics, assuming that they are reductive rather than eliminative, have got to hold that there are circumstances, specifiable without resort to semantical notions like truth, reference, correspondence or the like, such that, if a belief is formed *in* those circumstances, then it must be true. Verificationism adds to this only the idea that the circumstances are epistemic (they involve, for example, such idealizations as unrestricted access to the evidence) and that wouldn't seem to be the part that hurts. I guess what I'm saying is: if you're going to be a naturalist, there's no obvious reason not to be a verificationist. (And if you're *not* going to be a naturalist, why are you working on a causal theory of representation?)

The second point is this: verificationism isn't an ontological doctrine. It has usually, in the history of philosophy, been held with some sort of Idealistic malice aforethought, but that surely is an accident and one we can abstract from. The present sort of verificationism defines truth conditions by reference to the function of the cognitive mechanisms. Plausibly, the function of the cognitive mechanisms is to achieve, for the organism, epistemic access to the world. There is no reason on God's green earth why you shouldn't, in parsing that formula, construe "the world" Realistically.

Finally, verificationism isn't incompatible with a correspondance theory of truth. The teleology of the nervous system determines what must be the case if R represents S; and it follows from the analysis that if R represents S and the situation is teleologically normal, S must be true. This is because what R represents is its truth condition, and its truth condition is whatever causes its tokening in teleologically normal situations. But this is entirely compatible with holding that what *makes* R true in teleologically normal situations is that its truth condition obtains; that R corresponds, that is to say, to the way that the world is.

I see no way out of this: a causal theory must so characterize representation and normalcy that there is no misrepresentation in normal circumstances. My view is: if that is the price of a workable theory of representation, we ought simply to pay it.

NOTES

[1] Since we haven't any general and satisfactory way of saying which expressions *are* semantical (/intentional), it's left to intuition to determine when a formulation of C meets this condition. This will not, however, pose problems for the cases we will examine.

[2] I said that the formulation of naturalistic conditions for representation is *the least* that the vindication of an intentionalist psychology requires. What worries some philosophers is that there may be no *unique* answer to the question what something represents; e.g., that the representational content of a symbol (belief, etc.) may be *indeterminate* given the totality of physical fact. Notice that settling the question about naturalism doesn't automatically settle this question about determinacy. Even if it proves possible to give naturalistic necessary and sufficient conditions for representation, there might be more than one way to satisfy such conditions, hence more than one thing that R could be taken to represent. For purposes of the present paper, however, I propose to put questions about determinacy of representation entirely to one side and focus just on the prospects for naturalism.

[3] An example of the former: Propositional attitudes are relations to mental representations; mental representations are Ideas; Ideas are images; and Images represent what

they resemble. I take it that Hume held a view not entirely unlike this.

[4] In fact, Dretske gives the epistemic analysis as a condition upon '*R carries information about S*' rather than '*R* represents *S*'. This difference may *make* a difference and I'd have to attend to it if exposition were the goal. In much of what follows, however, I shall be less than sensitive to details of Dretske and Stampe's proposals. What I have in mind to exhibit are certain very pervasive characteristics of causal accounts; ones which I don't *think* can be avoided by tinkering.

[5] I should add that, though Stampe clearly thinks that you can, in principle, get representation without teleology, cases which turn on functional analysis loom large among his examples: "... one doubts whether statistical normality will get us far in dealing with living systems and with language or generally with matters of teleological nature. Here, I think we shall want to identify fidelity conditions with certain conditions of well functioning, of a functional system." (TCTLR, p. 51).

[6] Alternatively, you could go the disjunction route and say that the diameter of the coin represents some function of body temperature and air temperature. But this has the familiar consequence of rendering the covariance between *R* and *S* perfect and thus depriving us of examples of wild tokenings.

Dept. of Philosophy,
M.I.T.
Cambridge, MA 02139
U.S.A.