Chapter 5

Covariance II: Fodor

Jerry Fodor (1987) defends an account of the nature of mental representation that is remarkably similar to the one I have just discussed. The similarity is no accident; it will become clear as we go along that covariationists have a limited number of basic tools in the box. But there is no question that Fodor has added a few that are worth examining.

Background

Fodor begins by assuming the Representational Theory of Intentionality (which he calls the Representational Theory of Mind) and the Language-of-Thought Hypothesis (the hypothesis that mental representations are language-like symbols). Given these two assumptions, we can assume further that the problem of mental meaning generally reduces to the problem of understanding what it is for a primitive, nonlogical term of Mentalese to have a content. Given this focus, it will be convenient to have a convention for naming terms of Mentalese. In what follows, I shall write the term in Mentalese supposed to denote horses as |horse|; absolute values seem appropriate.

The basic idea—an idea Fodor calls the crude causal theory—is that symbol tokenings denote their causes and symbol types express the property whose instantiations reliably cause their tokenings.[1] Two problems immediately arise: that some noncats cause |cat|s, and that not all cats cause |cat|s.

The Disjunction Problem

Fodor calls the first problem the *disjunction problem*, for reasons that will emerge shortly. Suppose we try to describe a case of misrepresentation. A case of misrepresentation has to be a case like this: (1) Graycat causes a |dog| to occur in *S;* (2) |dog| expresses the property of being a dog in *S;* (3) Graycat is not a dog but a cat (of course). Now, since a cat (or, anyway, Graycat) causes a |dog| to occur in *S*, it follows that what |dog| must express in *S* is the property of being a dog-or-cat, or perhaps being a dog-or-Graycat, contrary to (2). It seems that any reason the crude causal theorist has to think that |dog| misrepresents Graycat as a dog is, for that theorist, a better reason to think that the content of |dog| has been misdescribed. Misrepresentation is always upstaged by a redescription of the alleged content. When the redescription is carried out, there is no misrepresentation. Hence, the crude causal theory implies that there is no misrepresentation.

It is tempting to reply that the causal route from Graycat to |dog| is not reliable. However, we can always make it reliable by describing the case in enough detail: There must be some situation in which Graycat reliably causes a |dog| in S—namely, the situation that obtained when, by hypothesis, Graycat was causally responsible for a |dog| in S. Moreover, there is such a thing as *systematic* misrepresentation: If I systematically misrepresent shrews as mice, this must be a case in which, according to the crude causal theory, shrews reliably cause |mouse|s in me. But there can't *be* such a case, since whatever is reliably caused by shrews is supposed to be a |shrew|.

Idealization

The obvious solution to the disjunction problem—one that Fodor himself briefly favored—is to idealize: In S |cat|s express the property of being a cat if, under ideal or optimal conditions, cats would reliably cause |cat|s in *S*. This move is, of course, familiar from our discussion of LOCKE, and it suffers from the same flaws: If the CTC is on the right track, there is no thoroughly naturalistic way to spell out the ideal conditions in question, and they won't eliminate error anyway.

Fodor in fact favors a different and far more ingenious solution.[2] The account pivots on the following claims: The fact that shrews sometimes cause |mouse|s in me depends on the fact that mice cause |mouse|s in me. On the other hand, the fact that mice cause |mouse|s in me doesn't depend on the fact that shrews sometimes cause |mouse|s in me. Mice look mousey to me, and that mousey look causes a |mouse|. But it is only because shrews also look mousey to me that shrews cause |mouse|s. Thus, if mice didn't cause |mouse|s, shrews wouldn't either. But it needn't work the other way; I could learn to distinguish shrews from mice, in which case mice would cause |mouse|s even though shrews would not.

This applies to the disjunction problem as follows; |mouse|s don't express the property of being a mouse-or-shrew, because the shrew-to-|mouse| connection is *asymmetrically dependent* on the mouse-to-|mouse| connection—the former connection would not exist but for the latter. In the case of genuinely disjunctive concepts, however, *A*-to-|*D*| connections are on a par with *B*-to-|*D*| connections, so |*D*|s express the property of being *A*-or-*B*.

Objections to Asymmetrical Dependence

I find this line unconvincing. Consider again the crucial counterfactuals:

(i) If mice didn't cause |mouse|s, shrews wouldn't cause |mouse|s.

(ii) If shrews didn't cause |mouse|s, mice wouldn't cause |mouse|s.

The alleged asymmetry depends on the claim that (i) is true and (ii) false. But is this really right? Shrews cause |mouse|s because they look like mice. Thus, if shrews didn't cause |mouse|s, that might be because (a) shrews didn't look like mice or because (b) mouse-looks didn't cause |mouse|s. If (b) were the culprit, though, mice wouldn't cause |mouse|s either, and that would make (ii) true.

It might seem that we can't blame (b) because the closest world in which shrews don't cause |mouse|s is the one in which (a) holds, not (b), since (b) requires a break in the rather central connection between mouse-looks and |mouse|s, whereas (a) requires only learning to distinguish mice and shrews. But this really isn't very persuasive. Perhaps shrews just look like mice to people, and finding out about shrews just makes them *uncertain* when they see either one. In a case like that, anything that will break the shrew-to-|mouse| connection will break the mouse-to-|mouse| connection as well. Even experts might perform randomly (perhaps the technology isn't adequate), even though they understand the difference perfectly well and can explain it to laypersons. Look what doctors do with diseases, or psychiatrists with psychoses and neuroses!

A variation on this theme suggests that the theory of asymmetrical dependence inverts the explanatory order: |mouse|s are wild when caused by shrews not because the more basic causal connection is with mice, but because |mouse|s express the property of being a mouse—something they might well do even if the dependence were symmetrical. Consider this story: In a certain tribe, all the youngsters are taught that they must catch a mouse for a certain potion the tribe needs. Mice are very rare, but only mice will do. Like all the other children, Broomhilda is taught how to catch a mouse (but not how to make the potion; only the medicine woman knows that). She is taught this by practicing on shrews. She has never seen a mouse, and she wouldn't recognize one if she saw one. Perhaps a mouse hasn't been seen in generations. Broomhilda knows there is a difference, however, for she knows at least this: Mice work in the potion, and shrews don't. Since the whole point of the training is to catch a mouse, the shrew-to-|*S*| connection (|*S*| is Broomhilda's internal representation) wouldn't exist but for the mouse-to-|*S*| connection. |*S*|s are, as Millikan (1984) would say, reproduced in Broomhilda because of the connection with mice. But also, given the way things are learned, the connection between |*S*|s and mice wouldn't exist if it were not for the connection between shrews and |*S*|s. There is no saying which connection is more fundamental. Hence, the asymmetrical-dependence doctrine must hold that |*S*| expresses the property of being a shrew-or-mouse. But it doesn't. It expresses the property of being a mouse, and *that* is why |*S*|s occasioned by shrews are wild.

A determined defender of asymmetrical dependence might avoid this criticism by claiming that scenarios like the ones just rehearsed that break down the asymmetry between (i) and (ii) are scenarios in which |mouse| (or |*S*|) is no longer a primitive term of Mentalese. But I don't think this will do, for it is pretty obvious that you can cook up similar scenarios for, say, |puce|.

Fodor's own reply is that the asymmetrical-dependence condition must apply *synchronically:* No matter how |mouse| and |shrew| are learned, current dispositions make the mouse-to-|mouse| connection primary. This strikes me as rather *ad hoc*, but let's see where it leads.

The picture Fodor has in mind is shown in figure 5.1. Mice cause mousey looks, which cause |mouse|s. Since shrews look mousey, they also cause mousey looks, thus poaching on the causal route from mice to |mouse|s and producing "wild" |mouse|s. Here are the relevant counterfactuals:

(1) If mice didn't cause |mouse|s, shrews wouldn't either. (T)

(2) If shrews didn't cause |mouse|s, mice wouldn't either. (F)

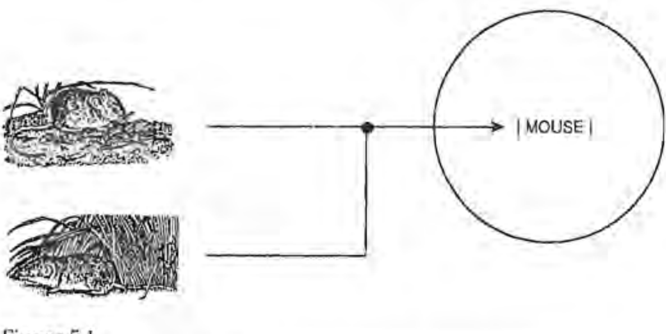

Figure 5.1
A shrew poaches on the mouse-to-|mouse| connection.

As indicated, we have the required asymmetrical dependence when (1) is true and (2) is false.

Start with (2). Figure 5.1 suggests two ways to break the shrew-to-|mouse| connection: (i) Mousey looks don't cause |mouse|s. But then mice won't cause |mouse|s either. Thus, (ii) shrews don't cause mousey looks. Perhaps shrews are extinct. More realistically, perhaps I come to *know something (perhaps tacitly) about how shrews and mice differ, and, as a result, shrews cease to even *look* like mice. But *mice* still look mousey, and hence they cause |mouse|s. So (2) is false, as desired.

Now consider (1). Again figure 5.1 suggests two ways to attack the mouse-to-|mouse| connection: (iii) Mousey looks don't cause |mouse|s. Since, by hypothesis, shrews poach by looking mousey, they will also cease to cause |mouse|s, and (1) is true, as required. Unfortunately, this way of making (1) true makes (2) true, as we just saw, so (iv) mice don't cause mousey looks. Perhaps mice become extinct, or acquire some disfiguring disease. But this won't affect the shrew-to-|mouse| connection, so (1) is false, contrary to requirements.

It looks as though the only way to have (1) true and (2) false is to employ different rules for evaluating them: Use (ii) to evaluate (2) and (iii) to evaluate (1). The possible worlds in which (1) is true and (2) is false are not the same possible worlds. To put it another way, there is no single interpretation that makes (1) true and (2) false. Therefore, (1) and (2) do not jointly express something about the connection between mice and shrews and that between mice and |mouse|s.

One might reply: "Well so what? All that means is that the definition of asymmetrical dependence is a bit messy. You have to say how (1) and (2) are to be (separately) evaluated." I wish I had a knock-down rebuttal to this reply, but I don't (even though I have the feeling there must be one). All I have is this: If you must get this tricky with the counterfactuals, you don't have a philosophical *explanation* any more; at best, you have a technically defensible equivalence between analysandum and analysans. It is hard to believe that the content of |*r*| is *mouse* rather than *mouse-or-shrew* BECAUSE

> if mice didn't cause |mouse|s because mousey looks didn't then shrews wouldn't cause |mouse|s either, and if shrews didn't cause |mouse|s becauseshrews didn't look mousey then mice would still cause |mouse|s (*ceteris paribus*, of course).

Maybe there *is* a way to make asymmetric dependence work without sacrificing explanation, but enough. As Fodor quite rightly says, the disjunction problem is the lesser of the two problems faced by the covariance theorist. Let's stop counting angels on pinheads and move on to where the action is.

Omniscience

> The Crude Causal Theory says, in effect, that a symbol expresses a property if it's nomologically necessary that *all* and *only* instances of the property cause tokenings of the symbol. (Fodor 1987, p. 100)

Lots of cats never cause |cat|s. (Well, to be safe, lots of rocks never cause |rock|s. But I prefer to stick with the cat-and-mouse game.) But why is that a problem for the covariationist? The crude causal theory was expressed thus: "... symbol tokenings denote their causes, and the symbol types express the property whose instantiations reliably cause their tokenings." It isn't obvious that *this* says, in effect, that *all* instantiations of the property cause tokenings of the symbol. Wherefore this strong and bothersome *all*? Granted, some cats don't cause |cat|s; but so what? Why isn't it enough that nothing else causes |cat|s (or, rather, that nothing else causes |cat|s in the basic way supposedly picked out by asymmetrical dependence)? It is well to get clear about this, because this seems to commit Fodor to the claim that cognitive systems are omniscient, and, as he admits, this is preposterous on its face. "... [P]roblems about the 'all' clause are, in my view," he writes, "very deep." So why is the 'all' clause *there*? Surprisingly, Fodor never answers this question, but the answer is quite simple: If some cats don't cause |*s*|s, then it seems that the extension of |*s*| should be the subset of cats that *do* cause |*s*|s. We need to rule out the possibility that |*s*| expresses the property of being, say, a black-and-white cat, or that of being Graycat. The only way the causal theorist can get around this is to insist on genuine covariation: All cats cause |*s*|s (or, anyway, any cat *would* cause an |*s*| *if given a fair chance*). But what is it to be *given a fair chance*?

The difficulty, of course, is that according to the CTC there is a fair chance that a cat will cause a |cat| only if the system is prepared to *attend* properly and to make the right *inferences* (or *inferences) on the basis of the right **knowledge*. But this sort of reply is clearly out of bounds; it will render the theory circular. Fodor realizes this but argues that, contrary to appearances, it *is* possible after all for a computationalist to specify causally sufficient conditions for a cat to cause a |cat|, or even for a proton to cause a |proton|, without trafficking in intentional or semantic notions.[3] Here is what he says:

> But though protons typically exert causal control over |proton|s via the activation of intentional mechanisms, a naturalistic semantics doesn't need to specify all that. All it needs is that the causal control should actually obtain, *however* it is mediated. The claim, to put it roughly but rather intuitively, is that it's sufficient for |proton| to express *proton* if there's a reliable correlation between protons and |proton|s, effected by a mechanism whose response is specific to psychophysical traces for which protons are *in fact* causally responsible. And that claim can be made in nonintentional, nonsemantic vocabulary. It just was.
>
> No doubt mechanisms that track nonobservables in the required way typically satisfy intentional characterizations (they're typically inferential) and semantic characterizations (they work because the inferences that they draw are sound). But that's OK because, on the one hand, the semantical/intentional properties of such mechanisms are, as it were, only contingently conditions for their success in tracking protons; and, on the other, what's required for |proton| to express *proton* is only that the tracking actually be successful. For purposes of semantic naturalization, *it's the existence of a reliable mind/world correlation that counts, not the mecha-*

> *nisms by which that correlation is effected.*[4] (Fodor 1987, pp. 121–122)

We have seen this move before; it is just the idea, scouted in the chapter 4 above, that the covariationist doesn't really owe us an account of the conditions under which, say, an *arbitrary* cat is guaranteed to produce a |cat|. All we need is (i) some guarantee that the relevant mechanism exists and (ii) a non-question-begging way to *pick out* that mechanism. The first part is plausible enough on general empirical grounds: There must be some circumstances in which cats are sufficient for |cat|s. And for Fodor the second is a cinch: "The mechanism that does the trick" does the trick! This is because all Fodor requires is a "naturalistic" way to pick out the mechanism, i.e., a way of picking out the mechanism without explicit use of intentional or semantic terms:

> What is required to relieve the worry that meaning will resist assimilation into the natural causal order 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 if* C where the vocabulary in which condition C is couched contains neither intentional nor semantical expressions. (Fodor 1984a, p. 2)

Fodor says that avoiding semantical and intentional expressions is only a *minimal* requirement, but in fact he takes it to be sufficient:

> The reference to 'mechanisms of belief fixation' perhaps makes this look circular, but it's not. At least not so far. Remember that we're assuming a functional theory of believing (though not, of course, a functional theory of believing that p; . . . .). On this assumption, having a belief is just being in a state with a certain causal role, so—in principle at least—we can pick out the belief states of an organism without resorting to semantical or intentional vocabulary. But then it follows that we can pick out the organism's mechanisms of belief fixation without recourse to semantical or intentional vocabulary: The mechanisms of belief *fixation*

> are, of course, the ones whose operations eventuate in the organism's having belief. (Fodor 1987, p. 105)

Perhaps we can pick out the mechanisms of belief fixation in "naturalistic" terms, but the CTC holds that we can't understand them or describe them without a healthy dose of representational lingo.

Well, admittedly, one philosopher's (or one scientist's) explanation is another's explanadum, but this seems like cheating to me. We are told that representation rests on a covariance between representation and representandum– between cats and |cats|s, for example. Covariance, in turn, is grounded in a mechanism that, under the right conditions, will produce a |cat| from a cat. According to the CTC, the mechanism in question can be understood only by appeal to inner representations, for the mechanism in question is one of inference from stored *knowledge. It follows that in order to understand the mechanism that the CTC invokes to explain the covariance between cats and |cat|s we must already understand representation and the explanatory role it plays in mental mechanisms. And that, by my lights, is enough to undermine the power of covariance theories to help us to understand the nature of representation in the CTC.

The problem, of course, is that it isn't enough to avoid intentional/semantic vocabulary; you must do it in a way that explains what representation is. It becomes obvious that just avoiding intentional/semantic vocabulary isn't enough when you see how easy it is. The problem, remember, was to say under what conditions cats are sufficient for |cat|s,[5] and to do it in naturalistic vocabulary. But look how easy it is: (i) Find an actual occasion in which a cat does cause a |cat|. Name that occasion *O*. (ii) Consider the mechanism that did the trick on occasion *O* (never mind how this worked, or whether it was peculiar to *O*), and call it *M*. (iii) Construct the desired counterfactual: Were *M* to operate on a cat in circumstances like those that obtained in *O*, a |cat| would result. Nothing to it!

The thing starts to come unraveled when we ask what *O* and *M* are like, for it is a fundamental consequence of the CTC that these must be *understood* inferentially (though, of course, they *can* be *picked out* naturalistically). The covariationist tells us that there is representation because there is covariance. The CTC tells us that there is covariance because there is representation, and Fodor agrees. But you can't have it both ways without undermining the explanatory power of one of the two doctrines. And since the philosophical problem before us is to explain representation in a way that will underwrite (not undermine) its explanatory role in the CTC, it is the covariationist doctrine that must go.

Here is a kind of analogy that may help clarify how I see the intellectual situation: Suppose someone tells you that the temperature of something depends on the amount of caloric in it. "What is caloric?" you ask. "Well," says your informant, "it is clear what one would like to say: Caloric is the stuff that increases in a thing when you raise its temperature. Of course, that's circular. But I can avoid the circle. Consider the mechanism that operates when you put tap water from the tap marked "C" in a pan on a lighted stove: Caloric is the stuff that mechanism causes to increase in the water." This identifies caloric without explaining it.

Idealization Again

We saw in chapter 4 that covariationists require idealization away from all sources of error. We are now in a position to put this point together with the point about circularity. The fact that you can't idealize away from error means that there is no *general* way to pick out a mechanism that will produce a |cat| in response to an arbitrary cat. Thus, the only way to do it is by reference to some *specific* instance or instances in which a cat *does* product a |cat|. We then say for all *S* that if *S* were in a situation like *that*, a cat would yield a |cat|. The sense that we no longer have an explanation of representation can be traced to the demonstrative. The account is essentially ostensive. "Representation," it says, "is when you have a case like *that*." Then you give an example or a sketch of what one would be like: "You know. It's like when you think there is a cat there because there *is* one there." There is no substantive way to specify the C in "In C, any cat would cause a cat in *S*," so the covariationist must, in the end, have recourse to ostension, and must hope you don't notice that there is no principled way to generalize on the example.

Chapter 6

Covariance III: Dretske

The Account in Knowledge and the Flow of Information

For present purposes, the account of the nature of representation as set out by Fred Dretske in his 1981 book *Knowledge and the Flow of Information* can be boiled down to the following two claims:

> (D1) The semantic content of a cognitive state *M* is a privileged part of its informational content, *viz.*, that informational content of *M* which is nested in no other informational content of *M*.[1]
>
> (D2) A cognitive state *M* of *O* has the proposition *p* as an informational content if the conditional probability that *p* is true, given that *O* is in *M*, is 1.

On this view, informational content is explicitly a matter of covariation between the representing state and the state represented. Indeed, Dretske often glosses D2 as the claim that *M* is a perfect indicator of the truth value of *p*. Perhaps it is worth emphasizing that, on this view, as on Fodor's and Locke's, *M*'s covariation with *p*'s holding isn't merely evidence that *M* has *p* as its informational content; it is constitutive: Representation *is* a special case of covariation on these accounts.

Misrepresentation

Notoriously, Dretske's account gives rise to difficulties in explaining the possibility of misrepresentation. It follows from D2 that if *p* is the informational content of *M*, then *p* is true. Hence, by D1, if *p* is the semantic content of *M*, *p* is true. It looks as if there can't be a false representation.

Dretske is alive to this difficulty, and he seeks to get around it in what should by now be the familiar way: idealization. The crux of his maneuver as it is set out in *Knowledge and the Flow of Information* is to distinguish the "learning situation," when conditions are supposed to be optimal, from ordinary situations, when they are not. In the former case, the occurrence of a token of *M* in the system is a perfect indicator that *p* is true. The system thus comes to rely on the occurrence of tokens of *M* to infer that *p* is the case. (Or perhaps, in simple systems, occurrences of tokens of *M* simply assume the control functions appropriate to *p*'s being true.) When conditions are not optimal, however, the indicator is no longer perfect: *O* can get into a token of state *M* even though *p* is not true. The inferential mechanism is still in place, however, so the organism infers that *p* is the case, contrary to fact.

The difficulties with this line of defense are well known.[2] First, only learned representations are covered, and a great deal is innate according to the CTC.[3] Second, there appears to be no noncircular way to distinguish the learning situation from others. On the face of it, organisms appear to learn to identify things without ever achieving perfection. I don't see how to get around this without simply stipulating that only situations in which an organization does develop a perfect indicator are to be counted as genuine learning situations. The danger of this move is that it runs a serious empirical risk: There is no reason to think there are any learning situations thus construed. Finally, it is hard to see how the occurrence of a token of *M* in *O* could be a perfect indicator that *p* is true if it is possible subsequently for a token of *M* to occur in *O* when *p* is false. What are we to say about what *would* have happened had one of these unfortunate circumstances obtained during the learning period? However this may be, it is certain that those tokens will not have *p* as an informational content and hence will not have *p* as a semantic content, so we are left without an account of misrepresentation.[4]

The fundamental source of these difficulties is the Lockean assumption that representation is essentially a matter of covariation. Since it is obvious that cognitive systems often misrepresent (i.e., often get into cognitive states that are not perfect indicators of the states of affairs they represent), cognitive representational content cannot be a species of informational content. The only way to save the idea in the face of this obvious fact is to attempt to define representational content in terms of informational content without making the former a species of the latter. There is really only one move that has a chance of working: The representational content of *M* is the informational content *M* would have under ideal conditions. And this is evidently the essence of Dretske's move (as it is of Fodor's and of the Lockean proto-theory discussed in chapter 4), except that Dretske implausibly holds that optimal conditions actually obtain during the "learning period." Idealization is forced on the covariationist by the obvious fact of misrepresentation, for misrepresentation is representation without covariation. Idealization is the only way to go with the idea that representation is covariation, for the covariationist, in the face of misrepresentation, *must* say, in effect, *"Well there would be covariation if things were nice."*

We have seen, however, that idealized covariance is problematic for the computationalist, for the CTC holds (i) that reliable mind-world covariation depends on representation, and not the other way around, and (ii) that it is not really possible to idealize away from error in any case. The theory of *Knowledge and the Flow of Information* doesn't help us with these fundamental problems.

Functional Meaning

Since the publication of *Knowledge and the Flow of Information*, Dretske has come up with what appears to be a different account of representation—an account specifically designed to deal with misrepresentation (Dretske 1986). This account identifies cognitive representation as a species of what Dretske calls functionally derived meaning:

> (M_f) *d*'s being *G* means$_f$ that *w* is $F =_{df}$ *d*'s function is to indicate the condition of *w*, and the way it performs this function is, in part, by indicating that *w* is *F* by its (*d*'s) being *G*.

In the 1986 work, Dretske claims to be primarily concerned with clarifying the appeal to functions in this analysis.[5] He emphasizes that the analysis itself is nothing particularly new and different. If that is correct, then the line of criticism I have been pressing against Lockean theories of representation should apply to this analysis regardless of how the appeal to functions is cashed out. Actually, a little work will reveal that Dretske's analysis in "Misrepresentation" is a slight variation on themes we have already rehearsed. The work is worth doing because it helps us to see how the constraints operating on covariationist theories always manage to push the advocates of covariation into the same basic configurations.

At first blush, M_f looks unpromising because of the use of the semantic term "indicate" on the right-hand side. One might well complain that if we knew what it was for a cognitive state to indicate something, we would already be home free. But this is premature, for Dretske actually has in mind the relatively innocent idea that *d*'s being *G* indicates that *w* is *F* just in case *d*'s being *G* covaries with *w*'s being *F*:

> (M_f) *d*'s being *G* means$_f$ that *w* is $F =_{df}$ *d*'s function is to covary with the condition of *w*, and the way it performs this function is, in part, by *d*'s being *G* when and only when *w* is *F*.

The appeal to functions in M_f does the same job it does in all Lockean accounts. Not all covariance is representation; sunburns don't represent overexposure to ultraviolet light, because it isn't a *function* of sunburns to covary with overexposure to ultraviolet light. If it is a *function* of *d*'s being *G* to covary with *w*'s being *F*, then we have representation (meaning$_f$).

This allows for misrepresentation because *d* can fail to perform its function. It is the function of a fuel gauge (let us suppose) to indicate the amount of fuel in the tank. It has this function even if the tank is full of water. When the tank is full of water, the gauge misrepresents the tank as full of fuel.[6]

Evaluating M_f

There are two ways to understand M_f. Compare the following:

(i) There is something *d* whose function is to covary with (indicate) the state of the world; a state of *d* represents *x* iff it covaries with *x* under ideal conditions.[7]

(ii) A state *M* represents *x* iff it is the function of *M* to covary with *x*.

Let us call the second variation the *specific-function variation,* to emphasize that in it each representation is identified via a specific function. In contrast, the first variation is a general-function variation, because it requires only a blanket function claim to the effect that there is a something *d* whose function is to indicate the state of the world.

We needn't trouble further with the general-function variation, since that evidently leads us over ground already explored. Does the specific-function variation give us a genuine alternative to the general-function variations already considered?

In Millikan's (1984) hands it does; the result will be the subject of the next chapter. But in Dretske's hands, the specific-function route returns us to familiar Lockean territory. The crucial point is this: On Dretske's view, it is a necessary condition of its being *R*'s function to covary with *x* that *R* would covary with *x* under normal (or optimal) conditions.[8] Idealized covariance is thus a necessary condition of meaning$_f$, and M_f thence inherits all the difficulties attendant on the idea that *x* represents *y* only if *x* would covary with *y* under ideal conditions.

Fixing Functions

I said above that Dretske is mainly concerned in "Misrepresentation" with the problem of clarifying the appeal to functions in M_f.[9] It is worth digressing to follow this line of thought because of what it reveals about the inner structure of the covariationist approach to representation. Here is the admirable illustration Dretske uses to introduce the problem:

> Some marine bacteria have internal magnets (called magnetosomes) that function like compass needles, aligning themselves (and, as a result, the bacteria) parallel to the earth's

> magnetic field. Since these magnetic lines incline downwards (toward geomagnetic north) in the northern hemisphere (upwards in the southern hemisphere), bacteria in the northern hemisphere, oriented by their magnetosomes, propel themselves toward geomagnetic north. The survival value of magnetotaxis (as the sensory mechanism is called) is not obvious, but it is reasonable to suppose that it functions so as to enable the bacteria to avoid surface water. Since these organisms are capable of living only in the absence of oxygen, movement towards geomagnetic north will take the bacteria away from oxygen-rich surface water and towards the comparatively oxygen-free sediment at the bottom. Southern hemispheric bacteria have their magnetosomes reversed, allowing them to swim toward geomagnetic south with the same beneficial results. Transplant a southern bacterium in the North Atlantic and it will destroy itself—swimming upwards (towards magnetic south) into the toxic, oxygen-rich surface water. (1986, p. 26)

According to M_f, if the orientation of the magnetosomes toward magnetic north is to mean$_f$ that oxygen-free water is in that direction, it must be the function of the magnetosomes to indicate the direction of oxygen-free water. The function clause in M_f is what identifies the representandum. But there seem to be several initially plausible ways to specify the function of the magnetosomes, and hence several initially plausible candidates for what is represented by the orientation of the magnetosomes. Hence are two choices:

> *Liberal* The function of the magnetosomes is to indicate the direction of oxygen-free water.
>
> *Conservative* The function of the magnetosomes is to indicate the direction of the surrounding magnetic field.

On the liberal reading, hemispherically transplanted bacteria are victims of misrepresentation; on the conservative reading they are not. On the conservative reading, even bar magnets don't fool them. Indeed, on the conservative reading, the only thing that could fool the bacterium would be a loss of polarity in the magnetosomes themselves, or some mechanical hindrance to their changing orientation. This raises the possibility that one can turn every case of misrepresentation into a case of the proper representation of something else simply by taking a more conservative view of the relevant functions.

In order to prevent this sort of deflationary trivialization of M_f, Dretske thinks he is obliged to find a way to rule out conservative construals of function in favor of liberal construals in every case in which misrepresentation is clearly possible. Only on the liberal reading can we say, for example, that in hemispherically transplanted bacteria the magnetosomes fail to perform their function—their function is to indicate the direction of oxygen-free water, they fail, and the organism destroys itself.

Dretske claims that a liberal reading is motivated only when the system exhibits a certain degree of complexity, a degree of complexity that magnetotaxic bacteria plausibly lack. The idea is relatively simple. Suppose we have two detection mechanisms that operate in parallel: the magnetosomes (as before) and a temperature sensor. Since surface water is generally warmer, an organism that prefers colder to warmer water will generally avoid oxygen-rich surface water. Imagine, further, some internal device *R* that changes the organism's direction of locomotion in response to either a change in the orientation of the magnetosomes or a change in the temperature sensor. The magnetosomes represent the direction of the magnetic field; the temperature sensor represents changes in temperature. What does *R* represent? According to Dretske, it represents the direction of oxygen-free water. No more proximal (conservative) representandum will do, according to Dretske, because the state of *R* never—even under optimal conditions—means_n anything less distal than something about the direction of oxygen-free water.

> Our problem with the bacteria was to find a way of having the orientation of its magnetosomes mean_f that oxygen-free water was in a certain direction without arbitrarily dismissing the possibility of its meaning_f that the magnetic field was aligned in that direction. We can now see that with the

> multiple resources described . . . this possibility can be non-arbitrarily dismissed. *R cannot* mean_f that [the temperature is changing] or [that the state of the temperature sensor is changing], because it doesn't, even under optimal conditions, mean_n this.[10] (1986, p. 34)

Even this will not be enough if, as Dretske points out, we are prepared to tolerate disjunctive meanings and say that *R* means_f that magnetosome orientation or temperature-sensor change has occurred. However, if the system can be classically conditioned, so that any proximal stimulus s_1 could come to substitute for (say) magnetosome orientation, then there is no definite disjunction of proximal stimuli to fall back upon. Throughout the system's conditioning history, different proximal stimuli will mediate the detection of *F*. "Therefore," Dretske writes,

> if we are to think of these cognitive mechanisms as having a time-invariant function at all (something that is implied by their continued—indeed, as a result of learning, more efficient—servicing of the associated need), then we *must* think of their function, not as indicating the nature of the proximal (even distal) conditions that trigger positive responses . . . but as indicating the condition *F* for which these diverse stimuli are signs. (1986, pp. 35–36)

This whole exercise is curious. Dretske is worried that misrepresentation will be ruled out by deflationary conservative function assignments. Thus, he needs to motivate

> A function of *F* is to indicate *x*

in cases in which *R* doesn't indicate *x*. The passage above makes it clear that Dretske accepts the following constraint on the relevant function assignments:

> A function of *R* is to indicate *x* only if *R* would covary with *x* under optimal conditions.

This is what does all the work in the arguments; deflationary conservative attributions of content are ruled out solely on the ground that the relevant covariance wouldn't hold "even under optimal conditions." The appeal to functions is completely idle here. It isn't that conservatives are wrong about *functions;* we can spell out their mistake—the mistake Dretske attributes to them, anyway—in the language of covariance without mentioning functions at all.

It is no surprise that, for Dretske, representation is where the covariance is. If you find covariance with a distal feature, not with a proximal one, then of course it is the distal feature that is represented. Dretske's point is that sufficiently complex systems can get into states that covary (ideally) with distal features but not with proximal ones, and hence that covariationists can deal with a deflationary conservative who tries to undermine the theory by systematically substituting correct representation of the more proximal for misrepresentation of the more distal.

Progress is progress, and one shouldn't knock it. Still, it is important to realize that blocking the deflationary conservative does nothing toward explaining idealized covariance in terms that do not beg the questions. Nor does it help with the disjunction problem, the problem that notoriously bedevils the account in *Knowledge and the Flow of Information*. That problem applies with full force to the doctrine of "Misrepresentation." Suppose that both mice and shrews cause (covary with) |*M*|s. Can |*M*|s be |mouse|s? That depends on whether a function of |*M*|s is to covary with mice but not with shrews. How are we to tell? Disappointingly, the only help "Misrepresentation" gives us with this question is to tell us how to use covariance to rule out function attributions. It is not a function of |*M*|s to covary with shrews if |*M*| wouldn't covary with shrews under ideal conditions. We are thus led right back into the familiar territory we have already explored.[11]

There is a glimmer of an idea here, though: Perhaps representation can be explained in terms of function, and functions can be explained without recourse to idealized covariance or to any other tacitly (or explicitly) intentional or semantic concepts. That is Millikan's strategy, the subject of the next chapter.

Chapter 7
Adaptational Role

Exposition

In an important series of publications, Ruth Millikan has offered a subtle and complex account of representation in terms of the adaptational role of symbols and the adaptational roles of the mechanisms that produce them and respond to ("interpret") them. (See especially Millikan 1984, 1986; see also Papineau 1984.) Millikan's treatment resists easy summary: I cannot, in the course of a single chapter, hope to do full justice to the theory. Instead, I will concentrate on the basics, illustrated in connection with what Millikan takes to be the most fundamental type of case. My purpose, as in my chapters on Fodor and Dretske, is the limited one of giving enough of the flavor of the theory to determine whether it is suitable as an explication of the concept of representation appealed to in the CTC.[1]

Here is the fundamental formulation:

> (M1) *C* is a truth condition for *r* in *S* $=_{df}$ *C*'s obtaining is a *basic factor* in a Normal Case for the performance of the Proper Functions of *r*- interpreters.[2]

Something *x* performs a Proper Function in a system *S* when it does the sort of thing the doing of which has been, historically, responsible for the replication of things of *x*'s type.[3] Thus, circulation of the blood is a Proper Function of hearts because it is the fact that hearts contribute to blood circulation in the way they do that has been, historically, responsible for the replication of hearts, and hence for the historical persistence of that type of organ.[4]

A basic factor[5] for the performance of a Proper Function is a factor that tracks the selective importance of the Normal Case of that function fulfilling its Proper Function.[6] A Normal Case is one in which the function is performed *successfully*—i.e., a case in which the item in question does its stuff and everything else conspires to produce the kind of result that is responsible for that item's replication.

Millikan's suggestive illustration is the bee dance. When a bee dance is performed properly, it has a certain *orientation*. A consequence of a properly performed bee dance is that spectator bees fly off in a direction that corresponds to the orientation of the dance. But the whole business is a *success* (i.e., comes off in the way that accounts for the replication of the relevant mechanisms) only if the bees find flowers in that direction (hence find pollen, hence produce nectar, hence honey, hence food, hence the means of survival). So a Normal Case—which need not be the *usual* case at all—is a case in which spectators respond to a dance by flying off in the direction that corresponds to the dance's orientation and find flowers as a result. Flowers' being in the relevant direction is, therefore, a *basic factor* for the dance interpreter's functioning Properly in response to a dance: no flowers, no pollen; no pollen, no food; no food, no reinforcement (via natural selection) for the mechanisms that respond to bee dances.

At first glance, it might seem that the Normal Case is loaded with *basic factors*: There has to be enough light, the wind can't be blowing too hard, the flowers can't be deadly to bees, there can't be bee predators or clouds of DDT in the relevant direction, and so on. What, one might wonder, makes it the case that the content of the bee dance is *flowers in the o-direction* (i.e., flowers in the direction corresponding to the orientation of the dance) rather than, say, *no predators in the o-direction*?

The answer is that the o-direction's being a predator-free zone is not part of the content of a bee dance because it is not basic to the Normal Case; rather, it is part of the background or boundary conditions of such cases.[7] In Normal Cases there are no predators, no nuclear wars, no hurricanes, no poisonous flowers, and so forth, and there is plenty of light and air, the earth moves in the usual way relative to the sun, etc. But although these are all necessary conditions for Normal interpreter performance, they are not basic factors; flower presence in the o-direction is the (or a) basic factor.

What makes the presence of flowers in the o-direction a basic factor, and the presence of light and the absence of predators a background condition? The idea is that it is the presence of flowers, and not the absence of predators, that tracks the selective significance of Normal Cases of bee-dance-interpreter function. Although this is clear enough intuitively, the notion of *basic factors* operative here has application outside of a selectionist framework. I will digress briefly to indicate what this application is because it has some independent interest, and because seeing it at work in a nonselectionist context helps to clarify what is going on in the selectionist contexts that are central to Millikan's account.

Basic Factors

Consider the simple pendulum law:

$$2\pi\sqrt{l/g} = T$$

This is an idealization; it captures only the contribution of length and gravity to period, ignoring friction and air resistance. It works because of two facts. First, length and gravity are independent of friction and air resistance; g and l don't change as a function of friction or air resistance. The idealization represented by equation 7.1 wouldn't work if increasing the friction shortened the length. Second, length and gravity are "basic factors" in this situation: One can begin with equation 7.1 and superimpose the effects of friction and air resistance, but one cannot begin with the contribution of friction and air resistance to period and then superimpose the contribution of length and gravity. *There is no such thing as the way a pendulum would behave were length not a factor, whereas there is such a thing as the way a pendulum would behave were air resistance and friction not factors.* Experimentally, the point is that it makes sense to reduce friction and air resistance and note whether the period beings to approximate

$2\pi\sqrt{l/g}$; we cannot manipulate length and note whether the period begins to approximate what it would be if friction, air resistance, and gravity were the only factors, for there is nothing to approximate.

The Normal Case for the functioning of the bee-dance interpreter is supposed to be an idealization in the same sense in which the simple pendulum law is an idealization: It focuses on the basic factors—the factors responsible for the basic phenomenon—and it ignores factors not constitutive of that basic phenomenon. In the Normal Case, the spectators react to the dance by flying off in the direction of the dance orientation. They find flowers. They gather pollen. They return to the hive and make honey. We don't mention the fact that clouds of DDT are not encountered, the fact that the wind doesn't blow them off course, that night doesn't fall, that the distance isn't so great that they drop from exhaustion, and so on, and so on. These are all relevant factors, but they aren't basic; they don't *ground* the phenomenon. There is an asymmetry between the role played by the flowers and that played by the absence of DDT. Although both are necessary conditions for success, the dance interpreter doesn't get replicated because of the cases in which spectators fail to encounter DDT, for there were presumably lots of such cases in which there were no flowers. Hence, "Spectators respond to the dance by flying off in the direction of the dance orientation, where they find no DDT and hence survive" does not describe a Normal Case in a way that specifies what makes Normal Cases of dance-interpreter function *Normal*.[8]

It is interesting to examine how Millikan's theory treats the case of the magnetotaxic bacteria discussed by Dretske. The problem, recall, is how to arbitrate between the liberal and conservative construals of the representational significance of magnetosome orientation. Liberals hold that the function of magnetosome orientation is to indicate the direction of safe water; conservatives hold that its function is to indicate the orientation of the magnetic field.

Millikan's theory comes down squarely on the side of liberalism, for the Normal Case is surely the case in which the locomotion mechanisms respond to magnetosome movement by propelling the organism away from the surface to safe (oxygen-free) water. Hence, the relevant *basic factor* is the presence of safe water in the o-direction (*viz.* down, toward geomagnetic north), for it is the role played by the interpreter—the locomotion mechanism, in this case—in getting the organism to safe water that is responsible for the replication of the interpreter (the locomotion mechanisms).

But why not say "The magnetosome indicates the direction of magnetic north, and the interpreter acts on the rule *Safe water in the direction of magnetic north*"? Because magnetic north's being in the o-direction is not a basic factor in the interpreter's (the locomotion system's) performing its function of getting the organism to safe water.[9] The interpreter was replicated because, in the Normal Case, it responded to magnetosome orientation in a way that led the organism down to safe water, not because it led the organism in the direction of magnetic north. From the point of view of selection, the fundamental connection is the one between magnetosome orientation and the direction of safe water, and it is the interpreter's role in mediating that connection that has led to to its replication.[10]

Evaluation

Despite the considerable merits of Millikan's account, it will not do as an account of the concept of representation required by the CTC. The reason is simple: The CTC is committed to an ahistorical notion of representation, and Millikan's notion is essentially historical.

Duplicates

Imagine a kind of duplicating machine that duplicates organisms—not by cloning, or by any other biochemical process that uses the information coded in the organism's DNA, but just as a copy machine duplicates a printed page without understanding it. The machine I have in mind produces a perfect physical duplicate of an organism without "understanding it." The "transporter" familiar to fans of *Star Trek* is presumably such a device, although it destroys the original in the process of producing a duplicate in another place.

Will a duplicate of an organism have the same representations as the original? The intuition is fairly widespread that it will. Indeed, the assumption behind the *Star Trek* transporter is that the duplicate is the same person who entered the transporter. There seems little doubt that, for the purposes of a psychology experiment, a molecule-by-molecule duplicate of a person would do as well as the original. To deny this seems to be to deny physicalism. (See the example attributed to Block by Fodor [1987, p. 40].)

The CTC goes farther than simple physicalism; it asserts that, in order to preserve the identity of a cognitive system (if not a whole mind or person), it suffices to produce a computational duplicate. Two systems running the same program on the same data structures are, according to the CTC, cognitively indistinguishable. Since physical duplicates are bound to be computational duplicates, the idea that physical duplicates must share cognitive states is a consequence of CTC. Computationalism holds that cognitive systems are automatic formal systems. It follows that any two automations of the same form system are the same cognitive system. However, a perfect physical duplicate of *D* will automate a formal system *F* if and only if *D* automates *F*.

An exactly parallel argument will work for any theory that holds that a cognitive system can be specified in a way that abstracts from its physical realization (e.g., connectionist theories), for to abstract from physical realization is to abstract from the history of those realizations. Current ahistorical state, according to this approach, determines current cognitive capacities, and hence must determine current representational content. The CTC abstracts from history for the same reason it abstracts from the actual items in a system's current environment: For better or worse, the CTC seeks a theory of cognitive capacities of the sort that might be brought to bear on radically different environments (with differing success, no doubt), and that might be realized in radically different stuff.

A conspicuous feature of Millikan's account is that perfect physical duplicates need not have the same representations. This is an immediate consequence of the fact that representational content is essentially a matter of *history* on Millikan's view. A perfect physical duplicate of me will have no evolutionary history or learning history at all. It will make no sense to ask of any of the mechanisms or subsystems of such a duplicate what factors were responsible for their replication: they weren't replicated in the sense required by Millikan's account.

It follows immediately that Millikan's account cannot be imported to explicate representation as that concept figures as an explanatory construct in computationalist theories of cognition. To put it crudely: According to computationalist accounts, history is an accidental property of a cognitive mechanism. According to computationalism, cognitive systems are individuated by their computational properties, and these are independent of history.

Someone with Millikanite leanings might be tempted to say that duplicate is cognitively equivalent to its original on the ground that it duplicates something with the right history. But this, in fact, amounts to abandoning the Millikan approach at a very fundamental level, for it amounts to conceding that it is really current state, ahistorically conceived, that underwrites cognitive content. What duplication produces, after all, is something that satisfies the same molecular blueprint as the original. If that is enough to ensure cognitive equivalence, then the significance of the original's history can only be that it resulted in an organism with the right molecular blueprint. History ceases to be of the essence, lapsing into the role of the only available technology for producing, without a prototype to copy, a system with the right molecular structure.

Another desperate attempt to somehow capitalize on Millikan's theory in order to give an account of representation that is consistent with computationalism is to think of the relevant history as somehow imminent in synchronically specified computational states. According to this idea, only when we think of the current state of a system as the product of a certain possible history does it make sense to characterize that system representationally. Although I think this idea has some merit of its own,[11] it is certainly no variation on Millikan, for it will surely be possible to describe various different histories that might have led to the same molecular blueprint or computational architecture—histories that will license quite different content attributions.[12] On Millikan's view, the point of appealing to actual history is to eliminate this kind of ambiguity (or indeterminacy, if you like); *actual* history determines the real content from among the possible ones.

A slightly more interesting possibility is that a physical duplicate would share the original's representations on the ground that selectionist/adaptationist explanations of the mechanisms in the original do carry over to the duplicate. The idea would be to claim, for example, that my duplicate's heart has the same Proper Functions as mine on the grounds that (i) the duplicate system is the way it is because *mine* is the way it is and (ii) my heart *does* have blood circulation as a Proper Function.

There is a good deal to be said for this suggestion (which I owe to Clayton Lewis). After all, genetic replication mechanisms are no more sensitive to selection history than the *Star Trek* transporter. Both mechanisms simply take things as they find them and replicate them, not on the basis of functional properties, but on the basis of physical properties on which functional properties supervene. It is hard to see, therefore, how to tell a story that will allow genetic replication to preserve Proper Functions but will not allow transporter replication to preserve them as well.[13]

Plausible as this line is, it won't help prop up a Millikanite reading of representation in the CTC. To repeat, the crucial issue is whether representation is grounded in the current state of the system, regardless of the history of that system. And on that issue the CTC is absolutely unambiguous: Computationally equivalent states are representationally equivalent. Yet computationally equivalent states need share no historical properties at all. From a computational perspective, historical properties are accidental properties. This is why computationalists can coherently hope to bypass the learning process. The CTC entails that if we *give* a system the same data structures that a natural system must acquire by learning, then (barring differences in computational architecture) we have a cognitively equivalent system. No computationalist can consistently suppose that a system that knows English must have *learned* it. And what goes for learning obviously goes for evolution as well. It is which data structures you have, not how you got them, that counts. Without this assumption, AI makes no sense at all.[14] No account that (like Millikan's) takes the history of a data structure seriously can be right for the CTC.

History and Belief

It is worth digressing a moment to point out that beliefs *are* individuated in a way that is sensitive to history. To borrow an example from Stich (1983), the belief my duplicate expresses with the words "I sold my car for a thousand dollars" is false because he didn't own the car in question, whereas the belief I express with the same words is true. We therefore have different beliefs, though by hypothesis we are computationally (indeed physically) equivalent. The problem is that my duplicate never acquired title to the car in question; I did. Hence, history matters to belief contents. If you are interested in belief contents, then you will do well to formulate an account that is sensitive to historical properties.

It doesn't follow from this, of course, that *representation* is sensitive to historical properties. In fact, nearly the opposite follows: Since data structures aren't sensitive to historical properties, it follows that beliefs aren't data structures. Moreover, it follows that beliefs don't inherit their contents from constituent data structures, as the RTI claims. This is one reason why philosophers who are interested in the semantics of belief (and of 'believes') are bound to misrepresent representation in the CTC: Data structures are insensitive to all sorts of things—such as historical properties—to which beliefs (and the other propositional attitudes) are exquisitely tuned. There is a good reason for this. As Stich points out, the CTC doesn't want to explain why my duplicate can't sell my car. Or, to put it as Ned Block does (in the example cited by Fodor, quoted above), some differences in belief are not legitimate sources of psychological variance.

Conclusion

Millikan's theory is sophisticated and complex. I have had to ignore most of the sophistication and complexity, for it lies aside from my main concern here, which is to discover what representation must be if the CTC is to be true and explanatory. The CTC entails that history is irrelevant to content, and Millikan's theory says that history is the very essence of the thing. This is no knock on Millikan; she was after intentionality (belief, etc.), not representation in the CTC.

The Evidential Value of Adaptational Role

Before going on to other possibilities, however, we should take note of a final point. When I introduced covariance theories in chapter 1, I suggested that one might be moved to ground representation in covariance because one discovers that a certain neural structure is an edge detector by noticing that it fires when and only when there is an edge in the organism's visual field. Obviously, an adequate treatment of representation should account for the evidential role of covariance. Should we also insist that an adequate treatment of representation account for the evidential role of selective history?

The problem with this suggestion is that we seldom *know* the selective history of anything. Or, rather, we seldom have knowledge of selective history that is epistemologically prior to the very facts about function and representation we are interested in establishing. We don't begin with the selective history of the bee-dance mechanisms and then infer the content of bee dances; we exploit covariance to infer a probable content and then *reconstruct* the selective history. Of course, the availability of a plausible reconstruction in a case of this kind is powerful corroboration. The discovery that bees find flowers when they fly off in response to bee dances, together with our knowledge that they *need* to find flowers, leaves us feeling satisfied that the content of the dance is *flowers in the direction of the dance orientation*. Nevertheless, if it always happened that spectator bees flew off in the direction of the dance orientation, found a pile of rocks, milled around, then went home, we would, I think, be justified in attributing the content *rocks in the direction of the dance orientation*, even though we would be mystified about the evolutionary significance of the whole business. It seems pretty clear that speculation about the evolutionary history (and even learning history) of central cognitive mechanisms will be possible only after we have a pretty good idea what representations are actually required. I don't see how we can hope to understand the adaptational significance of the abstract functional architecture of the brain until we know what cognitive capacities it underwrites. But to know that, we must traffic heavily in mental representation. Epistemologically, then, representation is prior to adaptational role. Moreover, and much more important, the explanatory order follows the epistemological order in this case. We can't explain the adaptational role of a cognitive capacity without presupposing mental representation, for a capacity is *cognitive* (and the particular cognitive capacity it is) only in virtue of its semantic characterization.

Those who hope to explain representation in terms of adaptational role, then, face a dilemma reminiscent of that faced by the Lockean covariance theorist. The adaptational significance of brain mechanisms is surely tracked by their cognitive significance to a large extent. If one wants to explain the adaptational significance of the brain, then, one must be in a position to specify the cognitive architecture of the brain. The adaptational significance of the brain presupposes its cognitive capacities, and (according to the CTC, at least) cognitive capacities rest on representational capacities. This order of explanation is undermined if representation is explained (defined) in terms of adaptational role. Hence the following dilemma: If one wants to explain (define) representation in terms of adaptational role, one cannot also explain the adaptational role of brain mechanisms in terms of its representational capacities, as the CTC proposes to do.