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:

Turns Dretske onto Ideal Conditions.

> (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.

Ok [illegible] [illegible] [illegible]

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 nonarbitrarily 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.

It was also Dretske's