CHAPTER 2

MISREPRESENTATION*

FRED DRETSKE

Epistemology is concerned with knowledge: how do we manage to get things right? There is a deeper question: how do we manage to get things wrong? How is it possible for physical systems to *misrepresent* the state of their surroundings?

The problem is not how, for example, a diagram, *d*, can misrepresent the world, *w*. For if we have another system, *r*, already possessed of representational powers, *d* can be used as an expressive extension of *r*, thereby participating in *r*'s representational successes and failures. When this occurs, *d* can come to mean that *w* is *F* when, in fact, *w* is not *F*, but *d*'s meaning derives, ultimately, from *r*. A chart depicting unemployment patterns over the past ten years can misrepresent this condition, but the chart's capacity for misrepresentation is derived from its role as an expressive instrument for agents, speakers of the language, who already have this power.

No, the problem is, rather, one of a system's powers of representation in so far as these powers do not derive from the representational efforts of another source. Unless we have some clue to how this is possible, we do not have a clue how naturally-evolving biological systems could have acquired the capacity for belief. For belief is, or so I shall assume, a *non-derived* representational capacity the exercise of which *can* yield a misrepresentation.

The capacity for misrepresentation is a part, perhaps only a small part, of the general problem of meaning or intentionality. Once we have meaning, we can, in our descriptions and explanations of human, animal, and perhaps even machine behaviour, lavish it on the systems we describe. Once we have intentionality, we can (to use Dennett's language) adopt the intentional stance [1] But what

* © Fred Dretske 1986

[1] D C Dennett, 'Intentional Systems', *Journal of Philosophy*, 68 (1971), 87–106, reprinted in *Brainstorms* (Montgomery, Vt., 1978)

(*besides* intentionality) gives us (and not, say, machines) the power to adopt this stance? Our ability to adopt this stance is an *expression*, not an analysis, of intentionality. The borrowed meaning of systems towards which we adopt appropriate attitudes tells us no more about the original capacity for misrepresentation than does a misplaced pin on a military map What we are after, so to speak, is *nature*'s way of making a mistake, the place where the misrepresentational buck stops. Only when we understand this shall we understand how grey matter can misrepresent the weather for tomorrow's picnic.

I. NATURAL SIGNS

Naturally-occurring signs mean something, and they do so without any assistance from us [2] Water does not flow uphill; hence, a northerly-flowing river means there is a downward gradient in that direction. Shadows to the east mean that the sun is in the west. A sudden force on the passengers in one direction means an acceleration of the train in the opposite direction. The power of these events or conditions to mean what they do is independent of the way we interpret them—or, indeed, of whether we interpret or recognize them at all. The dentist may *use* the X-ray to diagnose the condition of your upper right molar, but the dark shadows mean extensive decay has occurred whether or not he, or anyone else, appreciates this fact. Expanding metal indicates a rising temperature (and in this sense means that the temperature is rising) whether or not anyone, upon observing the former, comes to believe the latter. It meant that *before* intelligent organisms, capable of exploiting this fact (by building thermometers), inhabited the earth. If we are looking for the ultimate source of meaning, and with it an understanding of a system's power of misrepresentation, here, surely, is a promising place to begin.

Natural signs are indicators, more or less reliable indicators, and what they mean is what they indicate to be so. The power of a natural sign to mean something—for example, that Tommy has

[2] This needs some qualification, but it will do for the moment What a natural sign means often does depend on us, on what we *know* about relevant alternative possibilities or on how we *use* an associated device But if we don't know anything, or if the sign occurs in the operation of a device having no normal use, the sign still means something—just not, specifically, what we say it means under epistemically (or functionally) richer conditions I return to this point in n 8 below.

measles—is underwritten by certain objective constraints, certain lawful relations, between the sign (or the sign's having a certain property) and the condition that constitutes its meaning (Tommy's having measles). In most cases this relation is causal or lawful, one capable of supporting a counterfactual assertion to the effect that if the one condition had not obtained (if Tommy did not have measles), neither would the other (he would not have those red spots all over his face). Sometimes there are merely regularities, non-lawful but none the less pervasive, that help secure the connection between sign and significance It is partly the fact, presumably not itself lawful, that animals (for example, squirrels or woodpeckers) do not regularly ring doorbells while foraging for food that makes the ringing bell *mean* that someone (i e. some *person*) is at the door. If squirrels changed their habits (because, say, doorbells were made out of nuts), then a ringing doorbell would no longer mean what it now does But as things *now* stand, we can (usually) say that the bell would not be ringing unless someone was at the door, that the bell indicates someone's presence at the door, and that, therefore, that is what it means. But this subjunctively expressed dependency between the ringing bell and someone's presence at the door is a reflection of a regularity which, though not conventional, is not fully lawful either None the less, the doorbell retains its natural meaning as long as this regularity persists.

Beyond this I have nothing very systematic to say about what constitutes the natural meaning of an event or a condition.[3] I shall proceed with what I hope is a reasonably familiar notion, appealing (when necessary) to concrete examples. The project is to see how far one can go in understanding misrepresentation, the power of a condition (state, event, situation) r to mean (say, indicate) *falsely* that w is F (thereby misrepresenting w), in terms of a natural sign's meaning that w is F. Only when (or if) this project succeeds, or shows reasonable prospects of succeeding, will it, or might it, be necessary to look more carefully at what got smuggled in at the beginning.

Though natural meaning is a promising point of departure, it is hard to see how to get under way. Natural signs, though they mean something, though they can (in this sense) represent w (by indicating or meaning that w is F) are powerless to *misrepresent* anything

[3] I give a fuller account of it in F. Dretske, *Knowledge and the Flow of Information* (MIT Press, 1981), chs 1 and 2

Either they do their job right or they don't do it at all. The spots on Tommy's face certainly can mean that he has measles, but they mean this *only* when he has measles. If he doesn't have measles, then the spots don't mean this. Perhaps all they mean is that Tommy has been eating too many sweets.

Grice expresses this point by saying that an occurrence (a tokening of some natural sign) means (in what he calls the natural sense of 'meaning'—hereafter meaning_n) that P only if P.[4] He contrasts this sense of meaning with non-natural meaning where a sign can mean that P even though P is false. If we reserve the word 'meaning' (minus subscripts) for that species of meaning in which something can mean that w is F when w isn't F, the kind of meaning in which misrepresentation is possible, then meaning_n seems a poorly-qualified candidate for understanding meaning.

In speaking of signs and their natural meaning I should always be understood as referring to *particular* events, states or conditions: *this* track, *those* clouds, and *that* smoke. A sign type (for example, smoke) may be said to mean, in some natural sense, that there is fire even when every token of that type fails to mean_n this (because, occasionally, there is no fire). But this type-associated meaning, whatever its proper analysis, does *not* help us understand misrepresentation unless the individual tokens of that type *have* the type-associated meaning, unless particular puffs of smoke mean_n that there is fire when there is no fire. This, though, is not the case. A petrol gauge's registration of 'empty' (this *type* of event) can signify an empty tank, but when the tank is not empty, no particular registration of 'empty' by the gauge's pointer means_n that the tank is empty. Hence, no particular registration of the gauge misrepresents the amount of gas in the tank (by meaning_n that it is empty when it is not).

The inability of (particular) natural signs to misrepresent anything is sometimes obscured by the way we exploit them in manufactured devices. Interested as we may be in whether, and if so when, w becomes F, we concoct a device d whose various states are designed to function as natural signs of w's condition. Since this is how we use the device, we tend to say of some particular registration that d's being G (assuming this is the natural sign of w's being F) means that w is F even when, through malfunction or misuse, the system is

[4] P Grice, 'Meaning', *Philosophical Review*, 66 (1957), 377–88

failing to perform satisfactorily and w is not F. But this, clearly, is not what the particular pointer position means_n. This is what it is *supposed* to mean_n, what it was *designed* to mean_n, what (perhaps) tokens of type *normally* mean_n, but not what it *does* mean_n.

When there is a short circuit, the ring of the doorbell (regardless of what it was designed to indicate, regardless of what it normally indicates) does not indicate that the bellpush is being pressed. It still means_n (indicates) that there is electric current flowing in the doorbell circuit (one of the things it always meant_n), but the latter no longer means_n that the bellpush is being pressed. What the flow of current *now* means_n—and this is surely how we would judge it if we could *see* the bellpush, *see that* it was *not* being pressed—is that the system is malfunctioning or that there is a short circuit somewhere in the wiring. The *statement*, 'There is someone at the door', can mean that there is someone at the door even when no one is there, but the ringing doorbell cannot mean this when no one is there. Not, at least, if we are talking about meaning_n. If the bellpush is not being pressed, then we must look for something else for the ringing bell to mean_n. Often, we withdraw to some more proximal meaning_n, some condition or state of affairs in the normal chain of causal antecedents that *does* obtain (for example, the flow of current or the *cause* of the flow of current—for example, a short circuit) and designate it as the meaning_n of the ringing bell.

2. FUNCTIONAL MEANING

Granted, one may say, the doorbell's ringing cannot mean_n that someone is at the door when no one is there; still, in some related sense of meaning, it means this whether or not anyone is there. If this is not natural meaning (meaning_n), it is a close cousin.

Whether it is a cousin or not, there certainly is a kind of meaning that attaches to systems, or components of systems, for which there are identifiable *functions*. Consider, once again, the fuel gauge. It has a function: to pass along information about the amount of petrol in the tank. When things are working properly, the position of the needle is a natural sign of the contents of the tank. Its pointing to the left means_n that the tank is empty. Its pointing to the right means_n that the tank is full. And so on for the intermediate positions. But things sometimes go wrong: connections work loose, the battery goes dead, wires break. The gauge begins to register 'empty' when

the tank is still full. When this happens there is a tendency to say that the gauge misrepresents the contents of the tank. It *says* the tank is empty when it is not. It *means* (not, of course, $means_n$), but still means in *some* sense) that the tank is empty.

When *d*'s being *G* is, normally, a natural sign of *w*'s being *F*, when this is what it normally $means_n$, then there is a sense in which it means this whether or not *w* is *F if it is the function of d to indicate the condition of w*. Let us call this kind of meaning $meaning_f$—the subscript indicating that this is a functionally derived meaning.

(M_f) *d*'s being *G* $means_f$ that *w* is *F* = *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*

The position of the needle on the broken fuel gauge $means_f$ that the tank is empty because it is the gauge's function to indicate the amount of remaining fuel, and the way it performs this function is, in part, by indicating an empty tank when the gauge registers 'empty'.[5] And, for the same reason and in the same sense, the ringing doorbell says (i.e. $means_f$) that someone is at the door even when no one is there.

Whether or not M_f represents any progress in our attempt to naturalize meaning (and thus understand a system's non derivative power to misrepresent) depends on whether the functions in question can themselves be understood in some natural way. If these functions are (what I shall call) *assigned* functions, then $meaning_f$ is tainted with the purposes, intentions, and beliefs of those who assign the function from which $meaning_f$ derives its misrepresentational powers.[6] We shall not have tracked meaning, in so far as this involves the power of misrepresentation, to its original source. We shall merely have worked our way back, somewhat indirectly, to *our own* mysterious capacity for representation.

To understand what I mean by an *assigned* function, and the way

[5] I hope it is clear, that I am not here concerned with the word 'empty' (or the letter 'E') that might appear on the gauge. This symbol means empty whatever the gauge is doing, but this is purely conventional. I am concerned with what the pointer's position $means_n$ *whatever* we choose to print on the face of the instrument.

[6] L. Wright calls these 'conscious' functions; see his 'Functions', *Philosophical Review*, 82.2 (Apr. 1973), 142.

we (our intentions, purposes and beliefs) are implicated in a system's having such a function, consider the following case. A sensitive spring-operated scale, calibrated in fractions of a gram, is designed and used to determine the weight of very small objects. Unknown to both designers and users, the instrument is a sensitive indicator of altitude. By registering a reduced weight for things as altitude increases (note a things weight is a function of its height above sea level), the instrument *could* be used as a crude altimeter if the user attached a standard weight and noted the instrument's variable registration as altitude changed. Suppose, now, that under normal use in the laboratory the instrument malfunctions and registers 0.98 g. for an object weighing 1 g. Is it misrepresenting the *weight* of the object? Is it misrepresenting the *altitude* of the object? What does the reading of 0.98 g. mean? If we are talking about $meaning_n$, it clearly does not $mean_n$ that the object weighs 0.98 g. Nor does it $mean_n$ that the laboratory is 40,000 ft. above sea level. If we ask about $meaning_f$, though, it seems reasonable to say that the instrument's pointer says or indicates (i.e. $means_f$) that the object weighs 0.98 g. It is the function of this instrument to tell us what objects weigh, and it is telling us (incorrectly, as it turns out) that this object weighs 0.98 g.

But is the altitude being misrepresented? No. It should be noticed that the instrument cannot be misrepresenting *both* the altitude and the weight since a representation (or misrepresentation) of one presupposes a *fixity* (hence, *non*-representation) of the other.[7] Although the instrument *could* be used as an altimeter, it *is not* used that way. That is not its function. Its function is to register weight. That is the function we assign to it, the reason it was built and the explanation why it was built the way it was. Had our purposes been otherwise, it might have $meant_f$ something else. But they were not and it does not.

We sometimes change an instrument's assigned function. When we calibrate it, for example, we do not use it to measure what it is normally used to measure. Instead, we apply it to known quantities in order to use its indication as a (natural) sign of possible malfunction or inaccuracy in the instrument itself. In this case, a reading of 0.98 g. (for a weight *known* to be 1 g.) indicates that the spring has changed its characteristics, the pointer is bent, or some other

[7] A doorbell, for example, cannot $mean_n$ *both* that there is someone at the door *and* that there is a short circuit.

component is out of adjustment. We get a new functional meaning because our altered background knowledge (normally a result of different intentions and purposes) changes what the pointer's behaviour means$_n$. With *assigned* functions, the meanings$_f$ change as *our* purposes change.[8]

We sometimes use animals in the same way that we use instruments. Dogs have an acute sense of smell. Exploiting this fact, customs officers use dogs to detect concealed marijuana. When the dog wags its tail, barks, or does whatever it is trained to do when it smells marijuana, the dog's behaviour serves as a natural sign—a sign that the luggage contains marijuana. But this does not mean that the dog's behaviour (or the neural condition that triggers this behaviour) can misrepresent the contents of the luggage. The dog's behaviour may make the customs officer believe (falsely) that there is marijuana in the suitcase, but the dog's behaviour means$_f$ this only in a derived way If the dog is particularly good at its job, barking only when there is marijuana present, we can say that its bark indicates (i.e. means$_n$) that there is marijuana present. Furthermore, it means$_n$ this whether or not anyone interprets it as meaning$_n$ this, whether or not we *use* this natural sign for our own investigative purposes. But when there is no marijuana present, when the dog barks at an innocent box of herbs, the bark does *not* mean$_n$ that there is marijuana present. Nor does it mean$_f$ this in any sense that is independent of *our* interpretative activities. We can, of course, say what the bark means *to us* (that there is marijuana in the suitcase), but this way of talking merely reveals our own involvement in the meaning assigned to the dog's behaviour. *We* assign this meaning because this is the information we are *interested* in obtaining, the information we *expect* to get by using the dog in this way, the information the dog was trained to deliver. But if we set aside our interests and purposes, then, *when there is no marijuana present*, there is *no* sense in which the dog's bark means that there is

[8] It isn't the change of purpose *alone* that changes what something means$_n$ (hence, means$_f$). It is the fact that this change in use is accompanied by altered background knowledge, and meaning$_n$ changes as background knowledge changes. If, for example, *A* depends on both *B* and *C*, a changing *A* can mean$_n$ that *C* is changing *if* we know that *B* is constant. If we know that *C* is constant, it can mean$_n$ that *B* is changing. If we know nothing, it only means that either *B* or *C* is changing. Natural meaning is relative in this sense, but derelativizing it (by ignoring what we know and how we use a device) does not eliminate natural meaning It merely makes *less determinate* what things mean$_n$. For a fuller discussion of this point, see ch. 3 in Dretske, *Knowledge and the Flow of Information*.

marijuana in the suitcase. The only kind of misrepresentation occurring here is of the derived kind we are familiar with in maps, instruments, and language.

Therefore, if M$_f$ is to serve as a naturalized account of representation, where this is understood to include the power of *misrepresentation*, then the functions in question must be *natural* functions, functions a thing has which are independent of *our* interpretative intentions and purposes. What we are looking for are functions involving a system of natural signs that give these signs a content, and therefore a meaning (i.e. a meaning$_f$), that is not parasitic on the way *we* exploit them in our information-gathering activities, on the way we choose to interpret them.[9]

We need, then, some characterization of a system's natural functions. More particularly, since we are concerned with the function a system of natural signs might have, we are looking for what a sign is *supposed* to mean$_n$ where the 'supposed to' is cashed out in terms of the function of that sign (or sign system) in the organism's *own* cognitive economy. We want to know how *the dog* represents the contents of the luggage—what (if anything) the smell of the box means$_f$ *to it*.

3. NEEDS

The obvious place to look for natural functions is in biological systems having a variety of organs, mechanisms, and processes that were developed (flourished, preserved) *because* they played a vital information-gathering role in the species' adaptation to its surroundings. An information-gathering function, essential in most cases to the satisfaction of a biological need, can only be successfully realized in a system capable of occupying states that serve as natural signs of external (and sometimes *other* internal) conditions. If that cluster of photoreceptors we call the retina is to perform its function (whatever, exactly, we take this function to be), the various states of these receptors must mean$_n$ something about the character and distribution of one's optical surroundings. Just what the various

[9] I think much of our talk about the representational capacities of computers is of this assigned, hence derived, kind. It tells us nothing about the intrinsic power of a machine to represent or misrepresent anything Hence, nothing about the cognitive character of its internal states R Cummins, I think, gets it exactly right by distinguishing *cognition (a version of *assigned* meaning) from genuine cognition See his *Psychological Explanation* (MIT Press, 1983).

states these receptors mean_f will (in accordance with M_f) be determined by two things: (1) what it is the function of this receptor system to indicate, and (2) the meaning_n of the various states that enable the system to perform this function.

To illustrate the way M_f is supposed to work it is convenient to consider simple organisms with obvious biological needs—something or condition without which they could not survive. I say this is convenient because this approach to the problem of misrepresentation has its most compelling application to cognitive mechanisms subserving some basic biological need. And the consideration of *primitive* systems gives us the added advantage of avoiding that kind of circularity in the analysis that would be incurred by appealing to those kinds of 'needs' (for example, my need for a word processor) that are derived from desires (for example, my desire to produce faster, cleaner copy). We cannot bring desires in at this stage of the analysis since they already possess the kind of representational content that we are trying to understand.

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.[10] Since these magnetic lines incline downwards (towards geomagnetic north) in the northern hemisphere (upwards in the southern hemisphere), bacteria in the northern hemisphere, oriented by their internal magnetosomes, propel themselves towards geomagnetic north. The survival value of magnetotaxis (as this 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 towards 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.

If a bar magnet oriented in the opposite direction to the earth's magnetic field is held near these bacteria, they can be lured into a

[10] My source for this example is R. P. Blakemore and R. B. Frankel, 'Magnetic Navigation in Bacteria', *Scientific American*, 245 6 (Dec. 1981)

deadly environment. Although I shall return to the point in a moment (in order to question this line of reasoning), this appears to be a plausible instance of misrepresentation. Since, in the bacteria's normal habitat, the internal orientation of their magnetosomes means_n that there is relatively little oxygen in *that* direction, and since the organism needs precisely this piece of information in order to survive, it seems reasonable to say that it is the function of this sensory mechanism to serve the satisfaction of this need, to deliver this piece of information, to indicate that oxygen-free water is in *that* direction. If this is what it is *supposed* to mean_n, this is what it means_f. Hence, in the presence of the bar magnet and in accordance with M_f, the organism's sensory state misrepresents the location of oxygen-free water.

This is not to say, of course, that bacteria have *beliefs*, beliefs to the effect that there is little or no oxygen in *that* direction. The capacity for misrepresentation is only *one* dimension of intentionality, only *one* of the properties that a representational system must have to qualify as a belief system. To qualify as a belief, a representational content must also exhibit (among other things) the familiar opacity characteristic of the propositional attitudes, and, unless embellished in some way, meaning_f does not (yet) exhibit *this* level of intentionality. Our project, though, is more modest. We are looking for a naturalized form of misrepresentation and, if we do not yet have an account of false *belief*, we do, it seems, have a naturalized account of false *content*.

Apart from some terminological flourishes and a somewhat different way of structuring the problem, nothing I have said so far is particularly original. I have merely been retracing steps, some very significant steps, already taken by others. I am thinking especially of Stampe's seminal analysis of linguistic representation in which the (possibly false) content of a representation is identified with what would cause the representation to have the properties it has under conditions of well-functioning[11]; Enc's development of functional ideas to provide an account of the intentionality of cognitive states[12]; Fodor's application of teleological notions in

[11] D. Stampe, 'Toward a Causal Theory of Linguistic Representation', in P. French, T. Uehling, and H. Wettstein (edd.), *Midwest Studies in Philosophy*, Vol. 2 (University of Minnesota Press, 1977)

[12] B. Enc, 'Intentional States of Mechanical Devices', *Mind*, 91 (Apr. 1982), 362. Enc identified the content of a functional state with the (construction of the) properties of the event to which the system has the function of responding.

supplying a semantics for his 'language of thought'[13], and Millikan's powerful analysis of meaning in terms of the variety of proper functions a reproducible event (such as a sound or a gesture) might have.[14] I myself have tried to exploit (vaguely) functional ideas in my analysis of belief by defining a structure's semantic content in terms of the information it was developed to carry (hence, acquired the function of carrying).[15]

4. THE INDETERMINACY OF FUNCTION

Though this approach to the problem of meaning—and, hence, misrepresentation—has been explored in some depth, there remain obstacles to regarding it as even a promising sketch, let alone a finished portrait, of nature's way of making a mistake.

There is, first, the question of how to understand a system's ability to misrepresent something for which it has no biological need. If *O* does not need (or need to avoid) *F*, it cannot (on the present account) be the *natural* function of any of *O*'s cognitive systems to alert it to the presence (absence, location, approach, identity) of *F*. And without this, there is no possibility of *mis*representing something *as F*. Some internal state could still $mean_n$ that an *F* was present (in the way the state of Rover's detector system $means_n$ that the luggage contains marijuana), but this internal state cannot $mean_f$ this. What we have so far is a way of understanding how an organism might misrepresent the presence of food, an obstacle, a predator, or a mate (something there is a biological need to secure or avoid[16]), but no way of understanding how *we* can misrepresent things as, say, can-openers, tennis-rackets, tulips, or the jack of diamonds. Even if we suppose our nervous systems sophisticated enough to indicate (under normal conditions) the presence of such things, it surely cannot be the *natural* function of these neural states to signal the presence—much less, specific kinds—of kitchen utensils, sporting equipment, flowers, and playing cards.

[13] J Fodor, 'Psychosemantics, or Where Do Truth Conditions Come From?' *manuscript*

[14] R, Millikan, *Language, Thought and other Biological Categories* (MIT Press, 1984).

[15] Dretske. *Knowledge and the Flow of Information*, part 3

[16] Something for which there is, in Dennett's (earlier) language, an 'appropriate efferent continuation': see his *Content and Consciousness* (London, 1969)

I think this is a formidable, but *not* an insuperable, difficulty. For it seems clear that a cognitive system might develop so as to service, and hence have the natural function of servicing, some biological need without its representational (*and* misrepresentational) efforts being confined to these needs. In order to identify its natural predator, an organism might develop detectors of colour, shape, and movement of considerable discriminative power Equipped, then, with this capacity for differentiating various colours, shapes, and movements, the organism acquires, as a fringe benefit so to speak, the ability to identify (and, hence, misidentify) things for which it has no biological need. The creature may have no need for green leaves, but its need for pink blossoms has led to the development of a cognitive system whose various states are capable, because of their need-related $meaning_f$, to $mean_f$ that there are green leaves present. Perhaps, though having no need for such things, it has developed a taste for them and hence a way of representing them with elements that already have a $meaning_f$.

There is, however, a more serious objection to this approach to the problem of misrepresentation. Consider, once again, the bacteria. It was said that it was the function of their magnetotactic system to indicate the whereabouts of oxygen-free environments. But why describe the function of this system in this way? Why not say that it is the function of this system to indicate the direction of geomagnetic north? Perhaps, to be even more modest, we should assign to this sensor the function of indicating the whereabouts (direction) of magnetic (not necessarily *geo*magnetic) north. This primitive sensory mechanism is, after all, functioning perfectly well when, under the bar magnet's influence, it leads its possessor into a toxic environment. *Something* is going wrong in this case, of course, but I see no reason to place the blame on the sensory mechanism, no reason to say it is not performing *its* function. One may as well complain that a fuel gauge is not performing its function when the petrol tank is filled with water (and the driver is consequently misled about the amount of *petrol* he has left). Under such abnormal circumstances, the instrument is performing its duties in a perfectly satisfactory way—i.e., indicating the amount of liquid in the tank What has gone wrong is something for which the instrument itself is not responsible: namely, a breakdown in the normal correlations (between the quantity of liquid in the tank and the quantity of petrol in the tank) that make the gauge serviceable as a *fuel* gauge, that

allow it (when conditions are normal) to mean$_n$ that there is petrol in the tank. Similarly, there is nothing wrong with one's perceptual system when one consults a slow-running clock and is, as a result, misled about the time of day. It is the function of one's eyes to tell one what *the clock says*; it is the function of *the clock to* say what the time is. Getting things right about what you need to know is often a *shared* responsibility. You have to get G right and G has to get F right. Hence, even if it is F that you need, or need to know about, the function of the perceptual system may be only to inform you of G.

If we think about the bacterium's sensory system in this way, then *its* function is to align the organism with the prevailing magnetic field. It is, so to speak, the job of magnetic north to be the direction of oxygen-free water. By transplanting a northern bacterium in the southern hemisphere we can make things go awry, but *not* because a hemispheric transplant undergoes *sensory* disorientation. No, the magnetotactic system functions as it is supposed to function, as it was (presumably) evolved to function. The most that might be claimed is that there is some *cognitive* slip (the bacterium mistakenly 'infers' from its sensory condition that *that* is the direction of oxygen-free water). This sort of reply, however, begs the question by presupposing that the creature *already* has the conceptual or representational capacity to represent something *as* the direction of oxygen-free water. Our question is *whether* the organism has this capacity and, if so, where it comes from.[17]

Northern bacteria, it is true, have no need to live in northerly climes *qua* northerly climes. So to describe the function of the bacterium's detectors in terms of the role they play in identifying geomagnetic north is not to describe them in ways that reveal *how*

[17] Fodor (in a circulated draft of 'Why Paramecia Don't Have Mental Representations') distinguishes organisms for which a representational theory of mind is not appropriate (paramecia, for example) and ones for which it is (us, for example) in terms of the latter's ability to respond to non-nomic stimulus properties (properties that are not transducer-detectable). We, but not paramecia, are capable of representing something as, say, a crumpled shirt, and *being a crumpled shirt* is not a projectible property. In this article, Fodor is not concerned with the question of *where* we get this extraordinary representational power from (he suggests it requires inferential capacities). He is concerned only with offering it as a way of distinguishing us from a variety of other perceptual and quasi-cognitive systems.

I agree with Fodor about the importance and relevance of this distinction, but my present concern is to understand *how* a system could acquire the power to represent something in this way. The power to represent something *as* a crumpled shirt (where this implies the correlative ability to misrepresent it as such) is certainly not innate.

this function is related to the satisfaction of its needs. But we do not have to describe the function of a mechanism in terms of its possessor's ultimate biological needs.[18] It is the function of the heart to circulate the blood. Just *why* the blood needs to be circulated may be a mystery.

So the sticky question is: *given* that a system needs F, and *given* that mechanism M enables the organism to detect, identify or recognize F, *how* does the mechanism carry out this function? Does it do so by representing nearby Fs *as nearby* Fs or does it, perhaps, represent them merely *as nearby* Gs, trusting to nature (the correlation between F and G) for the satisfaction of its needs? To describe a cognitive mechanism as an F-detector (and, therefore, as a mechanism that plays a vital role in the satisfaction of an organism's needs) is not *yet* to tell the functional story by means of which this mechanism does its job. All we know when we know that O needs F and that m enables O to detect F is that M *either* means$_f$ that F is prsent *or* it means$_f$ that G is present where G is, in O's natural surroundings, a natural sign of F's presence (where G means$_n$ F).[19] If I need vitamin C, my perceptual–cognitive system should not automatically be credited with the capacity for recognizing objects *as* containing vitamin C (as meaning$_f$ that they contain vitamin C) just because it supplies me with the information required to satisfy

[18] Enc, 'Intentional States of Mechanical Devices', p. 168, says that a photoreceptor in the fruit-fly has the function of enabling the fly to reach humid spots (in virtue of the correlation between dark spots and humid spots). I have no objection to describing things in this way. But the question remains: *how* does it perform this function? We can answer this question without supposing that there is any mechanism of the fly whose function it is to indicate the degree of humidity. The sensory mechanism can perform this function if there is merely something to indicate the luminosity—i.e. a photoreceptor. *That* will enable the fly to reach humid spots. Likewise, the bacteria's magnetotactic sense *enables* (and, let us say, has the *function* of enabling) the bacteria to avoid oxygen-rich water. But the way it does it (it may be argued) is by having a sensor that indicates, and has the function of indicating, the direction of the magnetic field.

[19] In Fodor's way of putting the point (in 'Psychosemantics'), this is merely a way of saying that his identification of the semantics of M (some mental representation) with entry conditions (relative to a set of normalcy conditions) still leaves some slack. We can say that the entry condition is the absence (presence) of oxygen *or* a specific orientation of the magnetic field. Appeal to the selectional history of this mechanism won't decide *which* is the right specification of entry conditions—hence, won't tell us whether the bacteria are capable of *mis*representing anything. Fodor, I think, realizes this residual indeterminacy and makes the suggestive remark (n. 9) that this problem is an analogue of the problems of specifying the perceptual object for theories of perception.

this need. Representing things as oranges and lemons will do quite nicely.

The problem we face is the problem of accounting for the misrepresentational capacities of a system *without* doing so by artificially *inflating* the natural functions of such a system. We need some *principled* way of saying what the natural function of a mechanism is, what its various states not only $mean_n$, but what they $mean_f$. It sounds a bit far-fetched (to my ear at least) to describe the bacteria's sensory mechanism as indicating, and having the function of indicating, the whereabouts of oxygen. For this makes it sound as though it is not performing its function under deceptive conditions (for example, in the presence of a bar magnet). This is, after all, a *magneto*tactic, not a *chemo*tactic, sensor. But if we choose to describe the function of this sensor in this more modest way, we no longer have an example of a system with misrepresentational powers. A northern bacterium (transplanted in the southern hemisphere) will not be misrepresenting anything when, under the guidance of its magnetotactic sensor, it moves upwards (towards geomagnetic north) into the lethal surface water. The alignment of its magnetosomes will $mean_n$ what it has always $meant_n$, what it is its function to $mean_n$, what it is supposed to $mean_n$: namely, that *that* is the direction of magnetic north. The disaster can be blamed on the abnormal surroundings. Nor can we salvage some residual misrepresentational capacity by supposing that the bacterium, under the influence of a bar magnet, at least misrepresents the direction of geomagnetic north. For, once again, the same problem emerges: why suppose it is the function of this mechanism to indicate the direction of *geo*magnetic north rather than, simply, the direction of the surrounding magnetic field? If we describe the function only in the latter way, it becomes impossible to fool the organism, impossible to make it misrepresent anything. For its internal states only $mean_f$ that the magnetic field is pointing in *that* direction and (like a compass) this is always accurate.

5. FUNCTIONAL DETERMINATION

For the purpose of clarifying issues, I have confined the discussion to simple organisms with primitive representational capacities. It is not surprising, then, to find no clear and unambiguous capacity for misrepresentation at this level For this power—and, presumably,

the dependent capacity for belief—requires a certain threshold of complexity in the information-processing capabilities of a system. Somewhere between the single cell and man we cross that threshold. It is the purpose of this final section to describe the character of this threshold, to describe the *kind* of complexity responsible for the misrepresentational capabilities of higher organisms.

Suppose an organism (unlike our bacterium) has *two* ways of detecting the presence of some toxic substance F. This may be because the organism is equipped with two sense modalities, each (in their different way) sensitive to F (or some modally specific natural sign of F), or because a single sense modality exploits different external signs (or symptoms) of F. As an example of the latter, consider the way we might identify oak trees visually by either one of two ways: by the distinctive leaf pattern (in the summer) or by the characteristic texture and pattern of the bark (in winter). We have, then, two internal states or conditions, I_1 and I_2, each produced by a different chain of antecedent events, that are natural signs of the presence of F Each $means_n$ that F is present. Suppose, furthermore, that, having a need to escape from the toxic F, these internal states are harnessed to a third state, call it R, which triggers or releases a pattern of avoidance behaviour. Figure 2 assembles the relevant facts. R, of course, is also a natural sign of F. Under normal circumstances, R does not occur unless F is present. f_1 and f_2 are properties typical of normal Fs s_1 and s_2 are proximal stimuli.

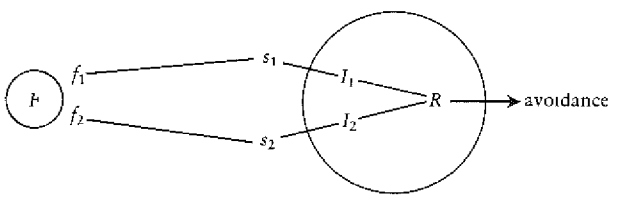

Fig. 2

If, now, we present the system with some ersatz F (analogous to the bar magnet with the bacteria), something exhibiting *some* of the properties of the real f (say f_1), we trigger a chain of events (s_1, I_1, R and avoidance) that normally occurs, and is really only appropriate, in the presence of F. If we look at the internal state R and ask what it

means$_f$ under these deceptive conditions, we find ourselves unable to say (as we could in the case of the bacteria) that it means$_f$ anything short of (i.e. more proximal than) F itself. Even though s_1 (by means of I_1) is triggering the occurrence of R, R does not mean$_n$ (hence, cannot mean$_f$) that s_1 (or f_1) is occurring. R is analogous to a light bulb connected to switches wired in parallel *either* of whose closure will turn the light on. When the bulb lights up, it does not mean$_n$ that switch no. 1 is closed even when it is this switch's closure that causes the light to go on. It does not mean$_n$ this, because there is no regular correlation between the bulb lighting up and switch no. 1 being closed (50 per cent of the time it is switch no. 2).

If we think of the detection system described above as having the function of enabling the organism to detect F, then the multiplicity of ways of detecting F has the consequence that certain internal states (for example, R) can indicate (hence mean$_f$) that F is present without indicating anything about the intermediate conditions (i.e. f_1 or s_1) that 'tell' it that F is present. 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 in Figure 2, this possibility can be *non*-arbitrarily dismissed. R *cannot* mean$_f$ that f_1 or s_1 is occurring, because it *does not*, even under optimal conditions, mean$_n$ this. We can therefore claim to have found a non-derivative case of misrepresentation (i.e., R's meaning$_f$ that F is present when it is not) which cannot be dismissed by redescribing what R means$_f$ so as to eliminate the appearance of misrepresentation. The threatened inflation of possible meanings$_f$, arising from the variety of ways a system's natural function might be described, has been blocked.

Still, it will be said, we *need not* accept this as a case of genuine misrepresentation *if* we are prepared to recognize that R has a *disjunctive* meaning$_n$. The lighting up of the bulb (connected to switches wired in parallel) does not mean$_n$ that any particular switch is on, but it does indicate that *one* of the switches is on. Similarly, it may be said, even though it is the function of the mechanism having R as its terminal state to alert the organism to the presence of F, it does so by R's indicating, and having the function of indicating, the occurrence of a certain disjunctive condition—namely, that either f_1 or f_2 (or s_1 or s_2). Our hypothetical organism mistakenly withdraws from F, *not* because it misrepresents the ersatz F as F, but because what it correctly indicates (i.e. that the ersatz f is either f_1 or f_2) is no longer correlated in the normal way with something's being F.

No matter how versatile a detection system we might design, no matter how many routes of informational access we might give an organism, the possibility will always exist of describing its function (and therefore the meaning$_f$ of its various states) as the detection of some highly disjunctive property of the proximal input. At least, this will always be possible *if* we have a determinate set of disjuncts to which we can retreat.

Suppose, however, that we have a system capable of some form of associative learning. Suppose, in other words, that through repeated exposures to *cs* (a conditioned stimulus) in the presence of F, a change takes place. R (and, hence, avoidance behaviour) can now be triggered by the occurrence of *cs* alone. Furthermore, it becomes clear that there is virtually no limit to the kind of stimulus that can acquire this 'displaced' effectiveness in triggering R and subsequent avoidance behaviour. Almost any s can become a *cs*, thereby assuming 'control' over R, by functioning (in the 'experience' of the organism) as a sign of F.

We now have a cognitive mechanism that not only transforms a variety of different sensory inputs (the s_i) into *one* output-determining state (R), but is capable of modifying the character of this many–one mapping over time. If we restrict ourselves to the sensory inputs (the s_i of Figure 2), R means$_n$ one thing at t_1 (for example, that either s_1 or s_2), something else at t_2 (for example, that either s_1 or s_2 or, through learning, cs_3), and something still different at a later time. Just *what* R means$_n$ will depend on the individual's learning history—on *what* s_i became cs_i *for it*. There is no *time-invariant* meaning$_n$ for R; hence, nothing that, through time, could be its function to indicate. In terms of the s_i that produce R, R can have no time-invariant meaning$_f$.

Of course, throughout this process, R continues to indicate the presence of F. It does so because, by hypothesis, any new s_i to which R becomes conditioned is a natural sign of F. Learning is a process in which stimuli that indicate the presence of F are, in their turn, indicated by some relevant internal state of the organism (R in this case). Therefore, 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 (the s_1 and f_1), but as indicating the condition (F) for which these diverse stimuli are signs. The mechanism just described has, then, as its natural function, the indication of the presence of F. Hence, the occurrence of R means$_f$ that F is present. It does not mean$_f$ that s_1 or s_2 or . . . s_x obtains, even though, at any given stage of development, it will mean$_n$ this for some definite value of x.

A system at this level of complexity, having not only multiple channels of access to what it needs to know about, but the resources for expanding its information-gathering resources, possesses, I submit, a genuine power of misrepresentation. When there is a breakdown in the normal chain of natural signs, when, say, cs_7 occurs (a learned sign of F) under circumstances in which it does not mean$_n$ that F is present (in the way that the broken clock does not mean$_n$ that it is 3.30 a.m.), R still means$_f$ (though not, of course, means$_n$) that F is present It means$_f$ this because that is what it is *supposed* to mean$_n$, what it is its natural function to mean$_n$, and there is available no other condition it can mean$_f$ [20]

[20] I am grateful to Berent Enc, Dennis Stampe, and Jerry Fodor for their helpful criticisms, both constructive and destructive, of earlier drafts of this essay