pathetically ill suited for fending for itself in any demanding environment. The barn swallow's fondness for carpentered nest sites might fool them into the view that it was some sort of pet, and whatever features of the cheetah convinced them that it was a creature of the wild might also be found in greyhounds and have been patiently encouraged by breeders. Artificial environments are themselves a part of nature, after all.

Prehistoric fiddling by intergalactic visitors with the DNA of earthly species cannot be ruled out, except on grounds that it is an entirely gratuitous fantasy. Nothing we have found (so far) on earth so much as hints that such a hypothesis is worth further exploration. (And note—I hasten to add, lest creationists take heart—that even if we were to discover and translate such a "trademark message" in our spare DNA, this would do nothing to rescind the claim of the theory of natural selection to explain all design in nature without invocation of a foresighted Designer-Creator *outside the system*. If the theory of evolution by natural selection can account for the existence of the people at NovaGene who dreamed up DNA branding, it can also account for the existence of any predecessors who may have left their signatures around for us to discover.) The power of the theory of natural selection is not the power to prove exactly how (pre-)history was, but only the power to prove how it could have been, given what we know about how things are.

Adaptationist thinking, then, may often be unable to answer particular questions about specific features of the historical mechanisms, the actual etiology, of a natural design development, even while it can succeed in formulating and even confirming—insofar as confirmation is ever possible—a functional analysis of the design. The difference between a design's having a free-floating (unrepresented) rationale in its ancestry and its having a represented rationale may well be indiscernible in the features of the design, but this uncertainty is independent of the confirmation of that rationale for that design. Moreover, as we shall see in the next chapter, the historical facts about the process of design development, even when we can discover them, are equally neutral when we move in the other direction: they are unable to resolve questions about the rationale of the design on which our interpretation of its activities depends. We should still hope science will eventually uncover the historical truth about these etiological details, but not because it will resolve all our Aristotelian "why" questions, even when they are cautiously and appropriately posed.

8 Evolution, Error, and Intentionality

Sometimes it takes years of debate for philosophers to discover what it is they really disagree about. Sometimes they talk past each other in long series of books and articles, never guessing at the root disagreement that divides them. But occasionally a day comes when something happens to coax the cat out of the bag. "Aha!" one philosopher exclaims to another, "so that's why you've been disagreeing with me, misunderstanding me, resisting my conclusions, puzzling me all these years!"

In the fall of 1985 I discovered what I took to be just such a submerged—perhaps even repressed—disagreement and guessed that it might take some shock tactics to push this embarrassing secret into the harsh glare of philosophical attention. There are few things more shocking to philosophers than strange bedfellows, so, in an earlier draft of this chapter which circulated widely in 1986, I drew up some deliberately oversimplified battle lines and picked sides—the good guys versus the bad guys. It worked. I was inundated with detailed, highly revealing responses from those I had challenged and from others who rose to the bait. By and large these reactions confirmed both my division of the field and my claims for its unacknowledged importance.

So constructive were the responses, however, even from those I had treated rather roughly—or misrepresented—in the earlier draft, that instead of just crowing "I told you so!" I should acknowledge at the outset that this heavily revised and expanded offspring of my

All but the last section of this chapter appears under the same title, in Y. Wilks and D. Partridge, eds., *Source Book on the Foundations of Artificial Intelligence* (Cambridge: Cambridge University Press, 1987), and is reprinted with permission.

earlier act of provocation owes a special debt to the comments of Tyler Burge, Fred Dretske, Jerry Fodor, John Haugeland, Saul Kripke, Ruth Millikan, Hilary Putnam, Richard Rorty, and Stephen Stich, and to many others, including especially Fred Adams, Peter Brown, Jerome Feldman, D. K. Modrak, Carolyn Ristau, Jonathan Schull, Stephen White, and Andrew Woodfield.

The Great Divide I want to display resists a simple, straightforward formulation, not surprisingly, but we can locate it by retracing the steps of my exploration, which began with a discovery about some philosophers' attitudes toward the interpretation of artifacts. The scales fell from my eyes during a discussion with Jerry Fodor and some other philosophers about a draft of a chapter of Fodor's *Psychosemantics* (1987). Scales often fall from my eyes when discussing things with Fodor, but this was the first time, so far as I can recall, that I actually found myself muttering "Aha!" under my breath. The chapter in question, "Meaning and the World Order," concerns Fred Dretske's attempts (1981, especially chapter 8; 1985; 1986) to solve the problem of misrepresentation. As an aid to understanding the issue, I had proposed to Fodor and the other participants in the discussion that we first discuss a dead simple case of misrepresentation: a coin-slot testing apparatus on a vending machine accepting a slug. "That sort of case is irrelevant," Fodor retorted instantly, "because after all, John Searle is right about one thing; he's right about artifacts like that. They don't have any intrinsic or original intentionality—only derived intentionality."

The doctrine of original intentionality is the claim that whereas some of our artifacts may have intentionality derived from us, we have original (or intrinsic) intentionality, utterly underived. Aristotle said that God is the Unmoved Mover, and this doctrine announces that we are Unmeant Meaners. I have never believed in it and have often argued against it. As Searle has noted, "Dennett . . . believes that nothing *literally* has any *intrinsic intentional* mental states" (1982, p. 57), and in the long-running debate between us (Searle 1980b, 1982, 1984, 1985; Dennett 1980b; Hofstadter and Dennett 1981; Dennett 1982c, 1984b, forthcoming f), I had assumed that Fodor was on my side on this particular point.

Did Fodor really believe that Searle is right about this? He said so. Dretske (1985) goes further, citing Searle's attack on artificial intelligence (Searle 1980) with approval, and drawing a sharp contrast between people and computers:

> I lack specialized skills, knowledge and understanding, but nothing that is essential to membership in the society of rational agents. With machines, though, and this includes the most sophisticated modern computers, it is different. They *do* lack something that is essential. (p. 23)

Others who have recently struggled with the problem of misrepresentation or error also seemed to me to fall on Searle's side of the fence: in particular, Tyler Burge (1986) and Saul Kripke (1982, especially p. 34ff). In fact, as we shall see, the problem of error impales all and only those who believe in original or intrinsic intentionality.

Are *original intentionality* and *intrinsic intentionality* the same thing? We will have to approach this question indirectly, by pursuing various attempts to draw a sharp distinction between the way our minds (or mental states) have meaning and the way other things do. We can begin with a familiar and intuitive distinction discussed by Haugeland. Our artifacts

> . . . only have meaning because we give it to them; their intentionality, like that of smoke signals and writing, is essentially borrowed, hence *derivative*. To put it bluntly: computers themselves don't mean anything by their tokens (any more than books do)—they only mean what we say they do. Genuine understanding, on the other hand, is intentional "in its own right" and not derivatively from something else. (1981, pp. 32–33)

Consider an encyclopedia. It has derived intentionality. It contains information about thousands of things in the world, but only insofar as it is a device designed and intended for our use. Suppose we "automate" our encyclopedia, putting all its data into a computer and turning its index into the basis for an elaborate question-answering system. No longer do we have to look up material in the volumes; we simply type in questions and receive answers. It might seem to naive users as if they were communicating with another person, another entity endowed with original intentionality, but we would know better. A question-answering system is still just a tool, and whatever meaning or aboutness we vest in it is just a by-product of our practices in using the device to serve our own goals. It has no goals of its own, except for the artificial and derived goal of "understanding" and "answering" our questions correctly.

But suppose we endow our computer with somewhat more autonomous, somewhat less slavish goals. For instance, a chess-playing computer has the (artificial, derived) goal of defeating its human opponent, of concealing what it "knows" from us, of tricking us perhaps. But still, surely, it is only our tool or toy, and although many of

its internal states have a sort of aboutness or intentionality—e.g., there are states that represent (and hence are about) the current board positions, and processes that investigate (and hence are about) various possible continuations of the game—this is just derived intentionality, not original intentionality.

This persuasive theme (it is not really an argument) has convinced more than a few thinkers that no artifact could have the sort of intentionality we have. Any computer program, any robot we might design and build, no matter how strong the illusion we may create that it has become a genuine agent, could never be a truly autonomous thinker with the same sort of original intentionality we enjoy. For the time being, let us suppose that this is the doctrine of original intentionality, and see where it leads.

The Case of the Wandering Two-Bitser

I will now press my vending machine example—the example Fodor insisted was irrelevant—explicitly, for it makes vivid exactly the points of disagreement and casts several recent controversies (about "individualistic psychology" and "narrow content," about error, about function) in a useful light. Consider a standard soft-drink vending machine, designed and built in the United States, and equipped with a transducer device for accepting and rejecting US quarters.[1] Let's call such a device a two-bitser. Normally, when a quarter is inserted into a two-bitser, the two-bitser goes into a state, call it *Q*, which "means" (note the scare-quotes) "I perceive/accept a genuine US quarter now." Such two-bitsers are quite clever and sophisticated, but hardly foolproof. They do "make mistakes" (more scare-quotes). That is, unmetaphorically, sometimes they go into state *Q* when a slug or other foreign object is inserted in them, and sometimes they reject perfectly legal quarters—they fail to go into state *Q* when they are *supposed to*. No doubt there are detectable patterns in the cases of "misperception." No doubt at least some of the cases of "misidentification" could be predicted by someone with enough knowledge of the relevant laws of physics and design parameters of the two-bitser's transducing machinery, so that it would be just as

1. This tactic is hardly novel. Among earlier discussions of intentionality drawing on such examples of simple discriminating mechanisms are MacKenzie, unpublished (1978), Ackermann 1972, and Enc 1982.

much a matter of physical law that objects of kind *K* would put the device into state *Q* as that quarters would. Objects of kind *K* would be good "slugs"—reliably "fooling" the transducer.

If objects of kind *K* became more common in the two-bitser's normal environment, we could expect the owners and designers of two-bitsers to develop more advanced and sensitive transducers that would reliably discriminate between genuine US quarters and slugs of kind *K*. Of course trickier counterfeits might then make their appearance, requiring further advances in the detecting transducers, and at some point such escalation of engineering would reach diminishing returns, for there is no such thing as a *foolproof* mechanism. In the meantime, the engineers and users are wise to make do with standard, rudimentary two-bitsers, since it is not cost effective to protect oneself against negligible abuses.

The only thing that makes the device a quarter-detector rather than a slug-detector or a quarter-*or*-slug-detector is the shared intention of the device's designers, builders, owners, users. It is only in the environment or context of those users and their intentions that we can single out some of the occasions of state *Q* as "veridical" and others as "mistaken." It is only relative to that context of intentions that we could justify calling the device a two-bitser in the first place.

I take it that so far I have Fodor, Searle, Dretske, Burge, Kripke, et al. nodding their agreement: that's just how it is with such artifacts; this is a textbook case of derived intentionality, laid bare. And so of course it embarrasses no one to admit that a particular two-bitser, straight from the American factory and with "Model A Two-Bitser" stamped right on it, might be installed on a Panamian soft-drink machine, where it proceeded to earn its keep as an accepter and rejecter of quarter-balboas, legal tender in Panama, and easily distinguished from US quarters by the design and writing stamped on them, but not by their weight, thickness, diameter, or material composition.

(I'm not making this up. I have it on excellent authority—Albert Erler of the Flying Eagle Shoppe, Rare Coins—that Panamanian quarter-balboas minted between 1966 and 1984 are indistinguishable from US quarters by standard vending machines. Small wonder, since they are struck from US quarter stock in American mints. And—to satisfy the curious, although it is strictly irrelevant to the example—the current official exchange rate for the quarter-balboa is indeed $.25!)

Such a two-bitser, whisked off to Panama (the poor man's Twin Earth), would still normally go into a certain physical state—the state with the physical features by which we used to identify state *Q*—whenever a US quarter or an object of kind *K* or a Panamanian quarter-balboa is inserted in it, but now a different set of such occasions count as the mistakes. In the new environment, US quarters count as slugs, as inducers of error, misperception, misrepresentation, just as much as objects of kind *K* do. After all, back in the United States a Panamanian quarter-balboa is a kind of slug.

Once our two-bitser is resident in Panama, should we say that the state we used to call *Q* still occurs? The physical state in which the device "accepts" coins still occurs, but should we now say that we should identify it as "realizing" a new state, *QB*, instead? Well, there is considerable freedom—not to say boredom—about what we should say, since after all a two-bitser is just an artifact, and talking about its perceptions and misperceptions, its veridical and nonveridical states—its intentionality, in short—is "just metaphor." The two-bitser's internal state, call it what you like, doesn't *really* (originally, intrinsically) mean either "US quarter here now" or "Panamanian quarter-balboa here now." It doesn't *really* mean anything. So Fodor, Searle, Dretske, Burge, and Kripke (*inter alia*) would insist.

The two-bitser was originally designed to be a detector of US quarters. That was its "proper function" (Millikan 1984), and, quite literally, its *raison d'être*. No one would have bothered bringing it into existence had not this purpose occurred to them. And given that this historical fact about its origin licenses a certain way of speaking, such a device may be primarily or originally characterized as a two-bitser, a thing whose function is to detect quarters, so that *relative to that function* we can identify both its veridical states and its errors.

This would not prevent a two-bitser from being wrested from its home niche and pressed into service with a new purpose—whatever new purpose the laws of physics certify it would reliably serve—as a *K*-detector, a quarter-balboa-detector, a doorstop, a deadly weapon. In its new role there might be a brief period of confusion or indeterminacy. How long a track record must something accumulate before it is no longer a two-bitser, but rather a quarter-balboa-detector (a q-balber)—or a doorstop or a deadly weapon? On its very debut as a q-balber, after ten years of faithful service as a two-bitser, is its state already a *veridical* detection of a quarter-balboa, or might there be a

sort of force-of-habit error of nostalgia, a mistaken identification of a quarter-balboa *as* a US quarter?

As described, the two-bitser differs strikingly from us in that it has no provision for memory of its past experiences—or even "memory" (in scare-quotes) for its past "experiences." But the latter, at least, could easily be provided, if it was thought to make a difference. To start with the simplest inroad into this topic, suppose the two-bitser (to refer to it by the name of its original baptism) is equipped with a counter, which after ten years of service stands at 1,435,792. Suppose it is not reset to zero during its flight to Panama, so that on its debut there the counter turns over to 1,435,793. Does this tip the balance in favor of the claim that it has not yet switched to the task of correctly identifying quarter-balboas? Would variations and complications on this theme drive your intuitions in different directions?

We can assure ourselves that nothing *intrinsic* about the two-bitser considered narrowly all by itself and independently of its prior history would distinguish it from a genuine q-balber, made to order on commission from the Panamanian government. Still, given its ancestry, is there not a problem about its function, its purpose, its meaning, on this first occasion when it goes into the state we are tempted to call *Q*? Is this a case of going into state *Q* (meaning "US quarter here now") or state *QB* (meaning "Panamanian quarter-balboa here now")? I would say, along with Millikan (1984), that whether its Panamanian debut counts as going into state *Q* or state *QB* depends on whether, in its new niche, it was *selected for* its capacity to detect quarter-balboas—literally selected, e.g., by the holder of the Panamanian Pepsi-Cola franchise. If it was so selected, then even though its new proprietors might have forgotten to reset its counter, its first "perceptual" act would count as a correct identification by a q-balber, for that is what it would *now* be *for*. (It would have acquired quarter-balboa detection as its proper function.) If, on the other hand, the two-bitser was sent to Panama by mistake, or if it arrived by sheer coincidence, its debut would mean nothing, though its utility might soon—immediately—be recognized and esteemed by the relevant authorities (those who could press it into service in a new role), and thereupon its *subsequent* states would count as tokens of *QB*.

Presumably Fodor et al. would be content to let me say this, since, after all, the two-bitser is just an artifact. It has no intrinsic, original intentionality, so there is no "deeper" fact of the matter we might try

to uncover. This is just a pragmatic matter of how best to talk, when talking metaphorically and anthropomorphically about the states of the device.

But we part company when I claim to apply precisely the same morals, the same pragmatic rules of interpretation, to the human case. In the case of human beings (at least), Fodor and company are sure that such deeper facts do exist—even if we cannot always find them. That is, they suppose that, independently of the power of any observer or interpreter to discover it, there is always a fact of the matter about what a person (or a person's mental state) *really means*. Now we might call their shared belief a belief in *intrinsic* intentionality, or perhaps even *objective* or *real* intentionality. There are differences among them about how to characterize, and name, this property of human minds, which I will continue to call *original intentionality*, but they all agree that minds are unlike the two-bitser in this regard, and this is what I now take to be the most fundamental point of disagreement between Fodor and me, between Searle and me, between Dretske and me, between Burge and me, etc. Once it was out in the open many things that had been puzzling me fell into place. At last I understood (and will shortly explain) why Fodor dislikes evolutionary hypotheses almost as much as he dislikes artificial intelligence (see, e.g., "Tom Swift and his Procedural Grandmother" in Fodor 1981a and the last chapter of Fodor 1983); why Dretske must go to such desperate lengths to give an account of error; why Burge's "anti-individualism" and Kripke's ruminations on rule-following, which strike some philosophers as deep and disturbing challenges to their complacency, have always struck me as great labors wasted in trying to break down an unlocked door.

I part company with these others because although they might agree with me (and Millikan) about what one should say in the case of the transported two-bitser, they say that we human beings are not just fancier, more sosphisticated two-bitsers. When we say that we go into the state of believing that we are perceiving a US quarter (or some genuine water as opposed to XYZ, or a genuine twinge of arthritis) this is no metaphor, no mere manner of speaking. A parallel example will sharpen the disagreement.

Suppose some human being, Jones, looks out the window and thereupon goes into the state of thinking he sees a horse. There may or may not be a horse out there for him to see, but the fact that he is in the mental state of thinking he sees a horse is not just a matter of

interpretation (these others say). Suppose the planet Twin Earth were just like Earth, save for having schmorses where we have horses. (Schmorses look for all the world like horses, and are well-nigh indistinguishable from horses by all but trained biologists with special apparatus, but they aren't horses, any more than dolphins are fish.) If we whisk Jones off to Twin Earth, land of the schmorses, and confront him in the relevant way with a schmorse, then either he really is, still, provoked into the state of believing he sees a horse (a mistaken, nonveridical belief) or he is provoked by that schmorse into believing, for the first time (and veridically), that he is seeing a schmorse. (For the sake of the example, let us suppose that Twin Earthians call schmorses *horses* (*chevaux, Pferde*, etc.) so that what Jones or a native Twin Earthian *says to himself*—or others—counts for nothing.) However hard it may be to determine exactly which state he is in, he is really in one or the other (or perhaps he really is in neither, so violently have we assaulted his cognitive system). Anyone who finds this intuition irresistible believes in original intentionality and has some distinguished company: Fodor, Searle, Dretske, Burge, and Kripke, but also Chisholm (1956, 1957), Nagel (1979, 1986), and Popper and Eccles (1977). Anyone who finds this intuition dubious if not downright dismissible can join me, the Churchlands (see especially Churchland and Churchland 1981), Davidson, Haugeland, Millikan, Rorty, Stalnaker, and our distinguished predecessors, Quine and Sellars, in the other corner (along with Douglas Hofstadter, Marvin Minsky, and almost everyone else in AI).

There, then, is a fairly major disagreement. Who is right? I cannot hope to refute the opposing tradition in the short compass of a chapter, but I will provide two different persuasions on behalf of my side: I will show what perplexities Fodor, Dretske, et al. entangle themselves in by clinging to their intuition, and I will provide a little thought experiment to motivate, if not substantiate, my rival view. First the thought experiment.

Designing a Robot

Suppose you decided, for whatever reasons, that you wanted to experience life in the twenty-fifth century, and suppose that the only known way of keeping your body alive that long required it to be placed in a hibernation device of sorts, where it would rest, slowed down and comatose, for as long as you liked. You could arrange to

climb into the support capsule, be put to sleep, and then automatically awakened and released in 2401. This is a time-honored science-fiction theme, of course.

Designing the capsule itself is not your only engineering problem, for the capsule must be protected and supplied with the requisite energy (for refrigeration or whatever) for over four hundred years. You will not be able to count on your children and grandchildren for this stewardship, of course, for they will be long dead before the year 2401, and you cannot presume that your more distant descendants, if any, will take a lively interest in your well-being. So you must design a supersystem to protect your capsule and to provide the energy it needs for four hundred years.

Here there are two basic strategies you might follow. On one, you should find the ideal location, as best you can foresee, for a fixed installation that will be well supplied with water, sunlight, and whatever else your capsule (and the supersystem itself) will need for the duration. The main drawback to such an installation or "plant" is that it cannot be moved if harm comes its way—if, say, someone decides to build a freeway right where it is located. The second alternative is much more sophisticated, but avoids this drawback: design a mobile facility to house your capsule along with the requisite sensors and early-warning devices so that it can move out of harm's way and seek out new energy sources as it needs them. In short, build a giant robot and install the capsule (with you inside) in it.

These two basic strategies are obviously copied from nature: they correspond roughly to the division between plants and animals. Since the latter, more sophisticated strategy better fits my purposes, we shall suppose that you decide to build a robot to house your capsule. You should try to design it so that above all else it "chooses" actions designed to further your best interests, of course. "Bad" moves and "wrong" turns are those that will tend to incapacitate it for the role of protecting you until 2401—which is its sole *raison d'être*. This is clearly a profoundly difficult engineering problem, calling for the highest level of expertise in designing a "vision" system to guide its locomotion, and other "sensory" and locomotory systems. And since you will be comtaose throughout and thus cannot stay awake to guide and plan its strategies, you will have to design it to generate its own plans in response to changing circumstances. It must "know" how to "seek out" and "recognize" and then exploit energy sources, how to move to safer territory, how to "anticipate" and then avoid dangers.

With so much to be done, and done fast, you had best rely whenever you can on economies: give your robot no more discriminatory prowess than it will probably need in order to distinguish what needs distinguishing in its world.

Your task will be made much more difficult by the fact that you cannot count on your robot being the only such robot around with such a mission. If your whim catches on, your robot may find itself competing with others (and with your human descendents) for limited supplies of energy, fresh water, lubricants, and the like. It would no doubt be wise to design it with enough sophistication in its control system to permit it to calculate the benefits and risks of cooperating with other robots, or of forming alliances for mutual benefit. (Any such calculation must be a "quick and dirty" approximation, arbitrarily truncated. See Dennett, forthcoming e.)

The result of this design project would be a robot capable of exhibiting self-control, since you must cede fine-grained real-time control to your artifact once you put yourself to sleep.[2] As such it will be capable of deriving its own subsidiary goals from its assessment of its current state and the import of that state for its ultimate goal (which is to preserve you). These secondary goals may take it far afield on century-long projects, some of which may be ill advised, in spite of your best efforts. Your robot may embark on actions antithetical to your purposes, even suicidal, having been convinced by another robot, perhaps, to subordinate its own life mission to some other.

But still, according to Fodor et al., this robot would have no original intentionality at all, but only the intentionality it derives from its artifactual role as your protector. Its simulacrum of mental states would be just that—not *real* deciding and seeing and wondering and planning, but only *as if* deciding and seeing and wondering and planning.

We should pause, for a moment, to make sure we understand what this claim encompasses. The imagined robot is certainly vastly more sophisticated than the humble two-bitser, and perhaps along the path to greater sophistication we have smuggled in some crucial new capacity that would vouchsafe the robot our kind of original intentionality. Note, for instance, that our imagined robot, to which we have granted the power to "plan" new courses of actions, to "learn" from past errors, to form allegiances, and to "communicate" with its com-

2. For more on control and self-control, see my *Elbow Room: The Varieties of Free Will Worth Wanting* (1984), chapter 3, "Control and Self-Control"; and forthcoming a.

petitors, would probably perform very creditably in any Turing Test to which we subjected it (see Dennett 1985a). Moreover, in order to do all this "planning" and "learning" and "communicating" it will almost certainly have to be provided with control structures that are rich in self-reflective, self-monitoring power, so that it will have a human-like access to its own internal states and be capable of reporting, avowing, and commenting upon what it "takes" to be the import of its own internal states. It will have "opinions" about what those states mean, and we should no doubt take those opinions seriously as very good evidence—probably the best evidence we can easily get—about what those states "mean" *metaphorically speaking* (remember: it's only an artifact). The two-bitser was given no such capacity to sway our interpretive judgments by issuing apparently confident "avowals."

There are several ways one might respond to this thought experiment, and we will explore the most promising in due course, but first I want to draw out the most striking implication of standing firm with our first intuition: no artifact, no matter how much AI wizardry is designed into it, has anything but derived intentionality. If we cling to this view, the conclusion forced upon us is that our own intentionality is exactly like that of the robot, for the science-fiction tale I have told is not new; it is just a variation on Dawkins's (1976) vision of us (and all other biological species) as "survival machines" designed to prolong the futures of our selfish genes. We are artifacts, in effect, designed over the eons as survival machines for genes that cannot act swiftly and informedly in their own interests. Our interests as we conceive them and the interests of our genes may well diverge—even though were it not for our genes' interests, we would not exist: their preservation is our original *raison d'être*, even if we can learn to ignore that goal and devise our own *summum bonum*, thanks to the intelligence our genes have installed in us. So our intentionality is derived from the intentionality of our "selfish" genes! *They* are the Unmeant Meaners, not us!

Reading Mother Nature's Mind

This vision of things, while it provides a satisfying answer to the question of whence came our own intentionality, does seem to leave us with an embarrassment, for it derives our own intentionality from entities—genes—whose intentionality is surely a paradigm case of

mere *as if* intentionality. How could the literal depend on the metaphorical? Moreover, there is surely this much disanalogy between my science-fiction tale and Dawkins's story: in my tale I supposed that there was conscious, deliberate, foresighted engineering involved in the creation of the robot, whereas even if we are, as Dawkins says, the product of a design process that has our genes as the primary beneficiary, that is a design process that utterly lacks a conscious, deliberate, foresighted engineer.

The chief beauty of the theory of natural selection is that it shows us how to eliminate this intelligent Artificer from our account of origins. And yet the process of natural selection is responsible for designs of great cunning. It is a bit outrageous to conceive of genes as clever designers; genes themselves could not be more stupid; *they* cannot reason or represent or figure out anything. They do not do the designing themselves; they are merely the beneficiaries of the design process. But then who or what does the designing? Mother Nature, of course, or more literally, the long, slow process of evolution by natural selection.

To me the most fascinating property of the process of evolution is its uncanny capacity to mirror *some* properties of the human mind (the intelligent Artificer) while being bereft of others. While it can never be stressed enough that natural selection operates with no foresight and no purpose, we should not lose sight of the fact that the process of natural selection has proven itself to be exquisitely sensitive to rationales, making myriads of discriminating "choices" and "recognizing" and "appreciating" many subtle relationships. To put it even more provocatively, when natural selection selects, it can "choose" a particular design *for one reason rather than another*, without ever consciously—or unconsciously!—"representing" either the choice or the reasons. (Hearts were chosen for their excellence as blood circulators, not for the captivating rhythm of their beating, though that *might* have been the reason something was "chosen" by natural selection.)

There is, I take it, no representation at all in the process of natural selection. And yet it certainly seems that we can give principled explanations of evolved design features that invoke, in effect, "what Mother Nature had in mind" when that feature was designed.[3]

3. "There must, after all, be a finite number of general principles that govern the activities of our various cognitive-state-making and cognitive-state-using mechanisms and there must be explanations of why these principles have historically worked to aid

Just as the Panamanian Pepsi-Cola franchise-holder can select the two-bitser *for* its talent at recognizing quarter-balboas, can adopt it *as* a quarter-balboa-detector, so evolution can select an organ *for* its capacity to oxygenate blood, can establish it *as* a lung. And it is only relative to just such design "choices" or evolution-"endorsed" purposes—*raisons d'être*—that we can identify behaviors, actions, perceptions, beliefs, or any of the other categories of folk psychology. (See Millikan 1984, 1986, for a forceful expression of this view.)

The idea that we are artifacts designed by natural selection is both compelling and familiar; some would go so far as to say that it is quite beyond serious controversy. Why, then, it is resisted not just by Creationists, but also (rather subliminally) by the likes of Fodor, Searle, Dretske, Burge, and Kripke? My hunch is because it has two rather unobvious implications that some find terribly unpalatable. First, if we are (just) artifacts, then what our innermost thoughts mean—and whether they mean anything at all—is something about which we, the very thinkers of those thoughts, have no special authority. The two-bitser turns into a q-balber without ever changing its inner nature; the state that used to mean one thing now means another. The same thing could in principle happen to us, if we are just artifacts, if our own intentionality is thus not original but derived. Those—such as Dretske and Burge—who have already renounced this traditional doctrine of privileged access can shrug off, or even welcome, that implication; it is the second implication that they resist: if we are such artifacts, not only have we no guaranteed privileged access to the deeper facts that fix the meanings of our thoughts, but *there are no such deeper facts.* Sometimes functional interpretation is obvious, but when it is not, when we go to read Mother Nature's mind, there is no text to be interpreted. When "the fact of the matter" about proper function is controversial—when more than one interpretation is well supported—there is no fact of the matter.

The tactic of treating evolution itself from the intentional stance needs further discussion and defense, but I want to approach the task indirectly. The issues will come into better focus, I think, if first we diagnose the resistance to this tactic—and its Siamese twin, the tactic of treating ourselves as artifacts—in recent work in philosophy of mind and language.

our survival. To suppose otherwise is to suppose that our cognitive life is an accidental epiphenomenal cloud hovering over mechanisms that *evolution devised with other things in mind.*" (Millikan 1986, p. 55; my emphasis)

Error, Disjunction, and Inflated Interpretation

Dretske's attempt (1981, 1985, 1986) to deal with these issues invokes a distinction between what he calls *natural meaning* and *functional meaning.* Natural meaning (*meaning*$_n$) is defined in such a way as to rule out misrepresentation; what a particular ringing of the doorbell means$_n$ depends on the integrity of the circuit that causes the ringing. "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 doorbutton is being depressed." "This is what is it *supposed* to mean$_n$, what it was *designed* to mean$_n$, what (perhaps) tokens of that type *normally* mean$_n$, but not what it *does* mean$_n$." (1986, p. 21)

It then falls to Dretske to define *functional meaning,* what it is for something to *mean*$_f$ that such-and-such, in such a way as to explain how a sign or state or event in some system can, on occasion, misrepresent something or "say" something false. But "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." (p. 22) Clearly, the meaning of the two-bitser's acceptance state *Q* is just such an assigned functional meaning, and Dretske would say of it: "That is the function we assign it, the reason it was built and the explanation for why it was built the way it was. Had our purposes been otherwise, it might have meant$_f$ something else." (p. 23)

Since merely *assigned* functional meaning is "tainted," Dretske must seek a further distinction. What he must characterize is the *natural* functions of the counterpart states of organisms, "functions a thing has which are independent of *our* interpretive intentions and purposes" (p. 25), so that he can then define natural functional meaning in terms of those functions.

> 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. (p. 25)

The obvious way to go, as we saw in the last section, is to substitute for our interpretive intentions and purposes the intentions and purposes of the organism's designer, Mother Nature—the process of natural selection—and ask ourselves what, in *that* scheme, any particular type of signal or state is designed to signal, supposed to mean.

Just as we would ultimately appeal to the engineers' rationales when deciding on the best account of representation and misrepresentation in our imagined survival-machine robot, so we can appeal to the discernible design rationales of natural selection in assigning content, and hence the power of *mis*representation, to event types in natural artifacts—organisms, ourselves included.

But although Dretske pays homage to those who have pursued that evolutionary path, and warily follows it some distance himself, he sees a problem. The problem is none other than the biological version of our question about what principled way there is to tell whether the state of the two-bitser (in some particular environment) means "quarter here now" or "quarter-balboa here now" or "thing of kind *F* or kind *G* or kind *K* here now." We must find an interpretation principle that assigns content, Dretske says, "without doing so by artificially *inflating* the natural functions of these systems"—while at the same time avoiding the too-deflationary principle which resolves all functional meaning into brute natural meaning, where misrepresentation is impossible.

Consider the classic case of what the frog's eye tells the frog's brain (Lettvin et al. 1959). Suppose we provoke a frog into catching and swallowing a lead pellet we toss at it (cf. Millikan 1986). If we interpret the signal coming from the eye as "telling" the frog that there is a fly flying toward it, then it is the eye that is passing mistaken information to the frog, whereas if we interpret that signal as merely signaling a dark moving patch on the retina, it is "telling the truth" and the error must be assigned to some later portion of the brain's processing (see Dennett 1969, p. 83). If we are strenuously minimal in our interpretations, the frog never makes a mistake, for every event in the relevant pathway in its nervous system can always be *de-interpreted* by adding disjunctions (the signal means something less demanding: fly *or* pellet *or* dark moving spot *or* slug of kind *K or* . . .) until we arrive back at the brute meaning$_n$ of the signal type, where misrepresentation is impossible. No matter how many layers of transducers contribute to a signal's specificity, there will always be a deflationary interpretation of its meaning as meaning$_n$ unless we relativize our account to some assumption of the normal (Normal, in Millikan's sense) function (see Dennett 1969, section 9, "Function and Content").

Dretske is worried about overendowing event types with content, attributing a more specific or sophisticated meaning to them than the facts dictate. But given the stinginess of Mother Nature the engineer, this otherwise laudable hermeneutical abstemiousness puts one at risk of failing to appreciate the "point," the real genius, of her inventions. A particularly instructive instance of the virtues of "inflationary" functional interpretation is Braitenberg's (1984) speculative answer to the question of why so many creatures—from fish to human beings—are equipped with special-purpose hardware that is wonderfully sensitive to visual patterns exhibiting symmetry around a vertical axis. There can be little doubt about what the deflationary description is of the content of these intricate transducers: they signal "instance of symmetry around vertical axis on the retina." But why? What is this for? The provision is so common that it must have a very general utility. Braitenberg asks what in the natural world (before there were church facades and suspension bridges) presents a vertically symmetrical view? Nothing in the plant world, and nothing in the terrain. Only this: other animals, *but only when they are facing the viewer!* (Rear views are often vertically symmetrical, but generally less strikingly so.) In other words, what a vertical-symmetry transducer tells one is (roughly) "someone is looking at you." Needless to say, this is typically a datum well worth an animal's attention, for the other creature, in whose cross-hairs the animal currently sits, may well be a predator—or a rival or a mate. And so it is not surprising that the normal effect of the symmetry detector's being turned ON is an immediate orientation reaction and (in the case of fish, for instance) preparation for flight. Is it inflationary to call this transducer a predator-detector? Or a predator-or-mate-or-rival-detector? If you were hired to design a fish's predator-detector, would you go for a more foolproof (but cumbersome, slow) transducer, or argue that this is really the very best sort of predator-detector to have, in which the false alarms are a small price to pay for its speed and its power to recognize relatively well-hidden predators?

Ecologically insignificant vertical symmetries count as *false* alarms only if we suppose the special-*purpose* wiring is *supposed* to "tell" the organism (roughly) "someone is looking at you." What *exactly* is the content of its deliverance? This quest for precision of content ascription, and for independence of interpretation, is the hallmark not only of Dretske's research program, but also of much of the theoretical work in philosophy of language and mind (the philosophical theory of meaning, broadly conceived). But at least in the case of the symmetry-detector (or whatever we want to call it) there is no "prin-

cipled" answer to that, beyond what we can support by appeal to the functions we can discover and make sense of in this way, in the normal operation of the transducer in nature.

We saw in the case of human-designed artifacts that we could use our appreciation of the costs and benefits of various design choices to upgrade our interpretation of the two-bitser's discriminatory talent from mere disk-of-weight-*w*-and-thickness-*t*-and diameter-*d*-and material-*m* detection to quarter detection (or quarter-balboa detection, depending on the user's intentions). This is, if you like, the fundamental tactic of artifact hermeneutics. Why should Dretske resist the same interpretive principle in the case of natural functional meaning? Because it is not "principled" enough, in his view. It would fail to satisfy our yearning for an account of what the natural event *really* means, what it means under the aspect of "original" or "intrinsic" intentionality.[4]

In "Machines and the Mental" (1985) Dretske claims that the fundamental difference between current computers and us is that while computers may process information by manipulating internal symbols of some sort, they have "no access, so to speak, to the *meaning* of these symbols, to the things the representations represent." (p. 26) This way of putting it suggests that Dretske is conflating two points: something's meaning something *to* or *for* a system or organism, and that system or organism's being in a position to know or recognize or intuit or introspect that fact from the inside.

4. Dretske happens to discuss the problem of predator detection in a passage that brings out this problem with his view: "If (certain) bacteria did not have something inside that meant that *that* was the direction of magnetic north, they could not orient themselves so as to avoid toxic surface water. They would perish. If, in other words, an animal's internal sensory states were not rich in information, intrinsic natural meaning, about the presence of prey, predators, cliffs, obstacles, water and heat, it could not survive." (1985, p. 29) The trouble is that, given Dretske's conservative demands on information, the symmetry-detector wouldn't count as sending a signal with information (intrinsic natural meaning) about predators but only about patterns of vertical symmetry on the retina, and while no doubt it could be, and normally would be, supplemented by further transducers designed to make finer-grained distinctions between predators, prey, mates, rivals, and members of ignorable species, these could be similarly crude in their actual discriminatory powers. If, as Dretske suggests, some bacteria can survive with only north-detectors (they don't need toxic-water-detectors, as it happens), other creatures can get by with mere symmetry-detectors, so the last sentence quoted above is just false: most animals survive and reproduce just fine without the benefit of states that are rich enough in (Dretskean) information to inform their owners about prey, predators, cliffs, and the like.

> Unless these symbols have what we might call an *intrinsic* [my emphasis] meaning, a meaning they possess which is independent of our communicative intentions and purposes, then this meaning *must* be irrelevant to assessing what the machine is doing when it manipulates them. (p. 28)

Dretske quite correctly insists that the meaning he is seeking for mental states must *make a real difference* in, and to, the life of the organism, but what he fails to see is that the meaning he seeks, while it is, in the case of an organism, independent of *our* intentions and purposes, is not independent of the intentions and purposes of Mother Nature, and hence is, in the end, just as derived and hence just as subject to indeterminacy of interpretation, as the meaning in our two-bitser.

Dretske attempts to escape this conclusion, and achieve "functional determination" in the face of threatened "functional indeterminacy," by devising a complicated story of how *learning* could make the crucial difference. According to Dretske, a learning organism can, through the process of repeated exposures to a variety of stimuli and the mechanism of associative learning, come to establish an internal state type that has a *definite, unique* function and hence functional meaning.

Confronted with our imagined robotic survival machine, Dretske's reaction is to suppose that in all likelihood some of its states do have natural (as opposed to merely assigned) functional meaning, in virtue of the learning history of the survival machine's first days or years of service. "I think we could (logically) create an artifact that *acquired* original intentionality, but not one that (at the moment of creation, as it were) *had* it" (personal correspondence). The functions dreamed of, and provided for, by its engineers are only *assigned* functions—however brilliantly the engineers anticipated the environment the machine ends up inhabiting—but once the machine has a chance to respond to the environment in a training or learning cycle, its states have at least the opportunity of acquiring natural (definite, unique) functional meaning—and not just the natural meaning in which misrepresentation is ruled out.

I will not present the details of this ingenious attempt because, for all its ingenuity, it won't work. Fodor (1987), in the chapter with which we began, shows why. First, it depends, as Fodor notes, on drawing a sharp line between the organism's learning period, when the internal state is developing its meaning, and the subsequent pe-

riod when its meaning is held to be fixed. Misrepresentation is possible, on Dretske's view, only in the second phase, but any line we draw must be arbitrary. (Does a whistle blow, Fodor wonders, signaling the end of the practice session and the beginning of playing for keeps?) Moreover, Fodor notes (not surprisingly), Dretske's account cannot provide for the fixed natural functional meaning of any innate, unlearned representative states.

Dretske does not view this as a shortcoming. So much the worse for innate concepts, he says. "I don't think there are, or can be, innate concepts or beliefs. . . . Beliefs and desires, *reasons* in general (the sort of thing covered by the intentional stance), are (or so I would like to argue) invoked to explain patterns of behavior that are acquired during the life history of the organism exhibiting the behavior (i.e., learned)" (personal correspondence).

The motivation for this stand can be brought out by considering an example. The first thing a baby cuckoo does when it hatches is to look around the nest for other eggs, its potential competitors for its adoptive parents' attention, and attempt to roll them over the edge. It surely has no inkling of the functional meaning of its activity, but that meaning is nevertheless there—*for* the organism and *to* the organism—unless we suppose by the latter phrase that the organism has to "have access" to that meaning, has to be in a position to reflect on it, or avow it, for instance. The rationale of the cuckoo's chillingly purposive activity is not in question; what remains to be investigated is to what extent the rationale is the fledgling's rationale and to what extent it is free-floating—merely what Mother Nature had in mind (see chapter 7). For Dretske, however, this is an all-or-nothing question, and it is tied to his intuition that there must be unique and unequivocal (natural functional) meanings for mental states.

Dretske seems to be trying to do two things at one stroke: first, he wants to draw a principled (and all-or-nothing) distinction between free-floating and—shall we say?—"fully appreciated" rationales; and second, he wants to remove all interpretive slack in the specification of the "actual" or "real" meaning of any such appreciated meaning-states. After all, if we appeal to our introspective intuitions, that is just how it seems: not only is there something we mean by our thoughts—something utterly determinate even if sometimes publicly ineffable—but it is our recognition or appreciation of *that meaning* that explains what we thereupon do. There certainly is a vast difference between the extremes represented by the fledgling cuckoo and, say,

the cool-headed and cold-blooded human murderer who "knows just what he is doing, and why," but Dretske wants to turn it into the wrong sort of difference. Echoing Searle, Dretske would sharply distinguish between syntax and semantics: in the human murderer, he would say, "it is the structure's having this meaning (its semantics), not just the structure that has this meaning (the syntax), which is relevant to explaining behavior" (personal correspondence; cf. Dretske 1985, p. 31). Even supposing Dretske could motivate the placement of such a threshold, dividing the spectrum of increasingly sophisticated cases into those where syntax does all the work and those where semantics comes unignorably into play, it is out of the question that the rigors of a learning history could break through *that* barrier, and somehow show an organism what its internal states "really meant."

Furthermore, *if* Dretske's learning-history move worked for learned representations, the very same move could work for innate representations "learned" by the organism's ancestors via natural selection over the eons. That is, after all, how we explain the advent of innate mechanisms—as arising out of a trial-and-error selection process over time. If, as Dretske supposes, "soft"-wiring can acquire natural functional meaning during an organism's lifetime, thanks to its relations to environmental events, "hard"-wiring could acquire the same natural functional meaning over the lifetime of the species.

And again, when do we blow the whistle and freeze, for all future time, the meaning of such a designed item? What started out as a two-bitser can become a q-balber; what started out as a wrist bone can become a panda's thumb (Gould 1980), and what started out as an innate representation meaning one thing to an organism can come, over time in a new environment, to mean something else to that organism's progeny. (There are further problems with Dretske's account, some well addressed by Fodor, but I will pass over them.)

What, then, does Fodor propose in place of Dretske's account? He too is exercised by the need for an account of how we can pin an error on an organism. ("No representation without misrepresentation" would be a good Fodorian motto.) And like Dretske, he draws the distinction between derivative and original intentionality:

> I'm prepared that it should turn out that smoke and tree rings represent only relative to our interests in predicting fires and ascertaining the ages of trees, that thermostats represent only relative to our interest in keeping the room warm, and that English words represent only relative to our intention to use

them to communicate our thoughts. I'm prepared, that is, that only mental states (hence, according to RTM [the Representional Theory of Mind], only mental representations) should turn out to have semantic properties *in the first instance;* hence, that a naturalized semantics should apply, strictu dictu, to mental representations only. (Fodor 1987, p. 99)

And then, like Dretske, he faces what he calls the disjunction problem. What principled or objective grounds can we have for saying the state means "quarter here now" (and hence is an error, when it occurs in perceptual response to a slug) instead of meaning "quarter *or* quarter-balboa *or* slug of kind *K or* . . ." (and hence, invariably, is not an error at all)? Fodor is no more immune than Dretske (or anyone else) to the fatal lure of teleology, of discovering what the relevant mechanism is "supposed to do," but he manfully resists:

I'm not sure that this teleology/optimality account is false, but I do find it thoroughly unsatisfying. . . . I think maybe we can get a theory of error without relying on notions of optimality or teleology; and if we can, we should. All else being equal, the less Pop-Darwinism the better, surely. (Fodor 1987, pp. 105–6)

I appreciate the candor with which Fodor expresses his discomfort with appeals to evolutionary hypotheses. (Elsewhere he finds he must help himself to a bit of "vulgar Darwinism" to buttress an account he needs of the functions of transducers.) Why, though, should he be so unwilling to follow down the path? Because he sees (I gather) that the most one can ever get from any such story, however well buttressed by scrupulously gathered facts from the fossil record, etc., is a story with all the potential for indeterminacy that we found in the tale of the transported two-bitser. And Fodor wants real, original, intrinsic meaning—not for the states of artifacts, heaven knows, for Searle is right about them!—but for our own mental representations.

Does Fodor have an account that will work better than Dretske's? No. His is equally ingenious, and equally forlorn. Suppose, Fodor says, "I see a cow which, stupidly, I misidentify. I take it, say, to be a horse. So taking it causes me to effect the tokening of a symbol; viz., I say 'horse'." There is an asymmetry, Fodor argues, between the causal relations that hold between horses and "horse" tokenings on the one hand and between cows and "horse" tokenings on the other:

In particular, misidentifying a cow as a horse wouldn't have led me to say 'horse' *except that there was independently a semantic relation between 'horse' tokenings and horses.* But for the fact that the word 'horse' expresses the property of *being a horse* (i.e., but for the fact that one calls *horses* 'horses'), it would not have been *that* word that taking a cow to be a horse would have caused me to utter. Whereas, by contrast, since 'horse' does mean *horse,* the fact that horses cause me to say 'horse' does not depend upon there being semantic—or, indeed, any—connection between 'horse' tokenings and cows. (Fodor 1987, pp. 107–8)

This doctrine of Fodor's then gets spelled out in terms of counterfactuals that hold under various circumstances. Again, without going into the details (for which see Akins, unpublished), let me just say that the trouble is that our nagging problem arises all over again. How does Fodor establish that, in his mental idiolect, "horse" means *horse*—and not *horse-or-other-quadruped-resembling-a-horse* (or something like that)? Either Fodor must go Searle's introspective route and declare that this is something he can just tell, from the inside, or he must appeal to the very sorts of design considerations, and the "teleology/optimality story" that he wants to resist. Those of us who have always loved to tell that story can only hope that he will come to acquire a taste for it, especially when he realizes how unpalatable and hard to swallow the alternatives are.

This brings me to Burge, who has also constructed a series of intuition pumps designed to reveal the truth to us about error. Burge has been arguing in a series of papers against a doctrine he calls *individualism,* a thesis about what facts settle questions about the content or meaning of an organism's mental states. According to individualism,

an individual's intentional states and events (types and tokens) could not be different from what they are, given the individual's physical, chemical, neural, or functional histories, where these histories are specified nonintentionally and in a way that is independent of physical or social conditions outside the individual's body. (1986, p. 4)

Or in other words:

The meaning or content of an individual's internal states could not be different from what it is, given the individual's *internal* history and constitution (considered independent of conditions outside its "body").

The falsehood of this thesis should not surprise us. After all, individualism is false of such simple items as two-bitsers. We changed the meaning of the two-bitser's internal state by simply moving it to Panama and giving it a new job to perform. Nothing structural or physical inside it changed, but the meaning of one of its states changed from *Q* to *QB* in virtue of its changed embedding in the world. In order to attribute meaning to functional states of an artifact,

you have to depend on assumptions about what it is supposed to do, and in order to get any leverage about that, you have to look to the wider world of purposes and prowesses. Burge's anti-individualistic thesis is then simply a special case of a very familiar observation: functional characterizations are relative not only to the embedding environment, but also to assumptions about optimality of design. (See, e.g., Wimsatt 1974. Burge seems to appreciate this in footnote 18 on p. 35.)

Moreover, Burge supports his anti-individualism with arguments that appeal to just the considerations that motivated our treatment of the two-bitser. For instance, he offers an extended argument (pp. 41ff) about a "person *P* who normally correctly perceives instances of a particular objective visible property *O*" by going into state *O'* and it turns out that in some circumstances, a different visible property, *C*, puts *P* into state *O'*. We can substitute "two-bitser" for "*P*", "*Q*" for "*O*'", "quarter" for "*O*", and "quarter-balboa" for "*C*", and notice that his argument is our old friend, without addition or omission.

But something is different: Burge leaves no room for indeterminacy of content; his formulations always presume that there is a fact of the matter about what something *precisely* means. And he makes it clear that he means to disassociate himself from the "stance-dependent" school of functional interpretation. He chooses to "ignore generalized arguments that mentalistic ascriptions are deeply indeterminate" (1986, p. 6) and announces his Realism by noting that psychology seems to presuppose the reality of beliefs and desires, and it seems to work. That is, psychology makes use of interpreted that-clauses, "—or what we might loosely call 'intentional content'." He adds, "I have seen no sound reason to believe that this use is merely heuristic, instrumentalistic, or second class in any other sense." (p. 8) That is why his thesis of anti-individualism seems so striking; he seems to be arguing for the remarkable view that *intrinsic* intentionality, *original* intentionality, is just as context sensitive as derived intentionality.

Although Burge, like Dretske and Fodor, is drawn inexorably to evolutionary considerations, he fails to see that his reliance on those very considerations must force him to give up his uncomplicated Realism about content. For instance, he champions Marr's (1982) theory of vision as a properly anti-individualistic instance of successful psychology without noticing that Marr's account is, like "engineering" accounts generally, dependent on strong (indeed too strong—

see Ramachandran, 1985a,b) optimality assumptions that depend on making sense of *what Mother Nature had in mind* for various subcomponents of the visual system. Without the tactic I have been calling artifact hermeneutics, Marr would be bereft of any principle for assigning content. Burge himself enunciates the upshot of the tactic:

> The methods of individuation and explanation are governed by the assumption that the subject has adapted to his or her environment sufficiently to obtain veridical information from it under certain normal conditions. If the properties and relations that *normally* caused visual impressions were regularly different from what they are, the individual would obtain different information and have visual experiences with different intentional content. (p. 35)

When we attribute content to some state or structure in Marr's model of vision, we must defend our attribution by claiming (in a paraphrase of Dretske on assigned functional meaning) that that is the function Mother Nature assigned this structure, the reason why it was built, and the explanation for why it was built the way it was. Had her purposes been otherwise, it might have meant$_f$ something else.

The method Burge endorses, then, must make the *methodological* assumption that the subject has adapted to his or her environment sufficiently so that when we come to assigning contents to the subject's states—when we adopt the intentional stance—the dictated attributions are those that come out veridical, *and useful*. Without the latter condition, Burge will be stuck with Fodor's and Dretske's problem of disjunctive dissipation of content, because you can always get veridicality at the expense of utility by adding disjuncts. Utility, however, is not an objective, determinate property, as the example of the two-bitser made clear. So contrary to what Burge assumes, he must relinquish the very feature that makes his conclusion so initially intriguing: his Realism about "intentional content," or in other words his belief that there is a variety of intrinsic or original intentionality that is not captured by our strategies for dealing with merely derived intentionality like that of the two-bitser.

The Realism about intentional content that Burge assumes, along with Fodor and the others, is also presupposed by Putnam, whose Twin Earth thought experiments (Putnam 1975a) set the agenda for much recent work on these issues. We can see this clearly, now, by contrasting our two-bitser with a Putnamian example. In the case of the two-bitser, the laws of nature do not suffice to single out what its

internal state *really means*—except on pain of making misrepresentation impossible. Relative to one rival interpretation or another, various of its moves count as errors, various of its states count as misrepresentations, but beyond the resources of artifact hermeneutics there are no deeper facts to settle disagreements.

Consider then the members of a Putnamian tribe who have a word, "glug," let us say, for the invisible, explosive gas they encounter in their marshes now and then. When we confront them with some acetylene, and they call it glug, are they making a mistake or not? All the gaseous hydrocarbon they have ever heretofore encountered, we can suppose, was methane, but they are unsophisticated about chemistry, so there is no ground to be discovered in their past behavior or current dispositions that would license a description of their glug-state as methane-detection *rather than* the more inclusive gaseous-hydrocarbon-detection. Presumably, gaseous hydrocarbon is a "natural kind" and so are its subspecies, acetylene, methane, propane, and their cousins. So the laws of nature will not suffice to favor one reading over the other. Is there a deeper fact of the matter, however, about what they *really mean* by "glug"? Of course once we educate them, they will have to *come* to mean one thing or the other by "glug," but in advance of these rather sweeping changes in their cognitive states, will there already be a fact about whether they believe the proposition that *there is methane present* or the proposition that *there is gaseous hydrocarbon present* when they express themselves by saying "Glug!"?

If, as seems likely, no answer can be wrung from exploitation of the intentional stance in their case, I would claim (along with Quine and the others on my side) that the meaning of their belief is simply indeterminate in this regard. It is not just that I can't tell, and they can't tell; there is nothing to tell. But Putnam, where he is being a Realist about intentional content (see chapter 10), would hold that there is a further fact, however inaccessible to us interpreters, that settles the questions about which cases of glug identification don't merely *count as* but *really are* errors, given what "glug" really means. Is this deeper fact any more accessible to the natives than to us outsiders? Realists divide on that question.

Burge and Dretske argue against the traditional doctrine of privileged access, and Searle and Fodor are at least extremely reluctant to acknowledge that their thinking ever rests on any appeal to such an outmoded idea. Kripke, however, is still willing to bring this

skeleton out of the closet. In Kripke's (1982) resurrection of Wittgenstein's puzzle about rule following, we find all our themes returning once more: a resistance to the machine analogy on grounds that meaning in machines is relative to "the intentions of the designer" (p. 34), and the immediately attendant problem of error:

> How is it determined when a malfunction occurs? . . . Depending on the intent of the designer, any particular phenomenon may or may not count as a machine malfunction. . . . Whether a machine ever malfunctions and, if so, when, is not a property of the machine itself as a physical object but is well defined only in terms of its program, as stipulated by its designer. (pp. 34–35)

This familiar declaration about the relativity and derivativeness of machine meaning is coupled with a frank unwillingness on Kripke's part to offer the same analysis in the case of human "malfunction." Why? Because it suggests that our own meaning would be as derivative, as inaccessible to us directly, as to any artifact:

> The idea that we lack "direct" access to the facts whether we mean plus or quus [Q or QB, in the two-bitser's case] is bizarre in any case. Do I not know, directly, and with a fair degree of certainty, that I mean plus? . . . There may be some facts about me to which my access is indirect, and about which I must form tentative hypotheses: but surely the fact as to what I mean by "plus" is not one of them! (p. 40)

This declaration is not necessarily Kripke speaking *in propria persona*, for it occurs in the midst of a dialectical response Kripke thinks Wittgenstein would make to a particular skeptical challenge, but he neglects to put any rebuttal in the mouth of the skeptic and is willing to acknowledge his sympathy for the position expressed.

And why not? Here, I think, we find as powerful and direct an expression as could be of the intuition that lies behind the belief in original intentionality. This is the doctrine Ruth Millikan calls *meaning rationalism*, and it is one of the central burdens of her important book, *Language, Thought, and Other Biological Categories*, to topple it from its traditional pedestal (Millikan 1984; see also Millikan, unpublished). Something has to give. Either you must abandon meaning rationalism—the idea that you are unlike the fledging cuckoo not only in having access, but also in having privileged access to your meanings—or you must abandon the naturalism that insists that you are, after all, just a product of natural selection, whose intentionality is thus derivative and hence potentially indeterminate.

Is Function in the Eye of the Beholder?

Attributions of intentional states to us cannot be sustained, I have claimed, without appeal to assumptions about "what Mother Nature had in mind," and now that we can see just how much weight that appeal must bear, it is high time to cash out the metaphor carefully.

Some have seen contradiction or at least an irresolvable tension, a symptom of deep theoretical incoherence, in my apparently willful use of anthropomorphic—more specifically, intentional—idioms to describe a process which I insist in the same breath to be mechanical, goalless, and lacking in foresight. Intentionality, according to Brentano, is supposed to be the "mark of the mental" and yet the chief beauty of the Darwinian theory is its elimination of Mind from the account of biological origins. What serious purpose could be served, then, by such a flagrantly deceptive metaphor? The same challenge could be put to Dawkins: How can it be wise to encourage people to think of natural selection as a watchmaker, while adding that this watchmaker is not only blind, but not even *trying* to make watches?

We can see more clearly the utility—in fact the inescapable utility—of the intentional stance in biology by looking at some other instances of its application. Genes are not the only micro-agents granted apparently mindful powers by sober biologists. Consider the following passages from L. Stryer's *Biochemistry* (1981) quoted by Alexander Rosenberg in "Intention and Action Among the Macromolecules" (1986b):

> A much more demanding *task* for these enzymes is to *discriminate* between similar amino acids. . . . However, the observed *error* frequency in vivo is only 1 in 3000, indicating that there must be subsequent *editing* steps to enhance fidelity. In fact the synthetase *corrects* its own *errors*. . . . How does the synthetase *avoid* hydrolyzing isoleucine-AMP, the *desired* intermediate? (pp. 664–65; Rosenberg's emphases)

It seems obvious that this is mere *as if* intentionality, a theorist's fiction, useful no doubt, but not to be taken seriously and literally. Macromolecules do not literally avoid anything or desire anything or discriminate anything. We, the interpreters or theorists, *make sense* of these processes by endowing them with mentalistic interpretations, but (one wants to say) the intentionality we attribute in these instances is neither real intrinsic intentionality, nor real derived intentionality, but mere *as if* intentionality.

The "cash value" of these metaphors, like the cash value of the metaphors about selfishness in genes that Dawkins scrupulously provides, is relatively close at hand. According to Rosenberg, "every state of a macromolecule which can be described in cognitive terms has both a unique, manageably long, purely physical characterization, and a unique, manageably describable disjunction of consequences" (p. 72), but this may be more an expression of an ideal that microbiologists firmly believe to be within their reach than an uncontroversial *fait accompli.* In similar fashion we could assure each other that for every vending machine known to exist, there is a unique, manageably long, manageably describable account of how it works, what would trick it, and why. That is, there are no mysteriously powerful coin detectors. Still, we can identify coin detectors as such—we can figure out that this is the competence that explains their existence—long before we know how to explain, mechanically, how that competence is achieved (or better: approximated).

Pending completion of our mechanical knowledge, we need the intentional characterizations of biology to keep track of what we are trying to explain, and even after we have all our mechanical explanations in place, we will continue to need the intentional level against which to measure the bargains Mother Nature has struck (see Dennett, forthcoming b).

This might be held sufficient methodological justification for the strategy of attributing intentional states to simple biological systems, but there is a further challenge to be considered. Rosenberg endorses the view—developed by many, but especially argued for in Dennett (1969 and 1983a)—that a defining mark of intentionality is failure of substitution ("intensionality") in the idioms that must be used to characterize the phenomena. He then notes that the biologists' attributions to macromolecules, selfish genes, and the like do not meet this condition; one can substitute ad lib without worry about a change in truth value, so long as the "subject" (the believer or desirer) is a gene or a macromolecule or some such simple mechanism. For instance, the proofreading enzyme does not recognize the error it corrects *qua* error. And it is not that the synthetase itself *desires* that isoleucine-AMP be the intermediate amino acid; it has no conception of isoleucine *qua* intermediate.

The disappearance of intensionality at the macromolecular level at first seems a telling objection to the persistent use of intentional idioms to characterize that level, but if we leave it at that we miss a

still deeper level at which the missing intensionality reappears. The synthetase may not desire that isoleucine-AMP be the intermediate amino acid, but it is only *qua* intermediate that the isoleucine is "desired" at all—as an unsubstitutable part in a design whose rationale is "appreciated" by the process of natural selection itself. And while the proofreading enzyme has no inkling that it is correcting errors *qua* errors, Mother Nature does! That is, it is only *qua* error that the items thus eliminated provoked the creation of the "proofreading" competence of the enzymes in the first place. The enzyme itself is just one of Nature's lowly soldiers, "theirs not to reason why, theirs but to do or die," but *there is* a reason why they do what they do, a reason "recognized" by natural selection itself.

Is there a reason, really, why these enzymes do what they do? Some biologists, peering into the abyss that has just opened, are tempted to renounce *all* talk of function and purpose, and they are right about one thing: there is no stable intermediate position.[5] If you are prepared to make any claims about the function of biological entities—for instance, if you want to maintain that it is perfectly respectable to say that eyes are for seeing and the eagle's wings for flying—then you take on a commitment to the principle that natural *selection* is well named. In Sober's (1984) terms, there is not just selection *of* features but selection *for* features. If you proceed to assert such claims, you find that they resist substitution in the classical manner of intentional contexts. Just as George IV wondered whether Scott was the author of *Waverley* without wondering whether Scott was Scott, so natural selection "desired" that isoleucine be the intermediate without desiring that isoleucine be isoleucine. And without this "discriminating" prowess of natural selection, we would not be able to sustain functional interpretations at all.

Certainly we can describe all processes of natural selection without appeal to such intentional language, but at enormous cost of cumbersomeness, lack of generality, and unwanted detail. We would miss the pattern that was there, the pattern that permits prediction and

5. Rosenberg (1986b):

> Among evolutionary biologists, there are those who condemn the identification of anatomical structures as having specific adaptational significance, on the ground that such structures do not face selection individually, but only in the company of the rest of the organism. This makes ascriptions of adaptational "content" to a part of the organism indeterminate, since a different ascription together with other adjustments in our adaptational identifications can result in the same level of fitness for the whole organism. In the philosophy of psychology, the dual of this thesis is reflected in the indeterminacy of interpretation.

supports counterfactuals. The "why" questions we can ask about the engineering of our robot, which have answers that allude to the conscious, deliberate, explicit reasonings of the engineers (in most cases) have their parallels when the topic is organisms and their "engineering." If we work out the rationales of these bits of organic genius, we will be left having to attribute—but not in any mysterious way—an emergent appreciation or recognition of those rationales to natural selection itself.

How can natural selection do this without intelligence? It does not consciously seek out these rationales, but when it stumbles on them, the brute requirements of replication ensure that it "recognizes" their value. The illusion of intelligence is created because of our limited perspective on the process; evolution may well have tried all the "stupid moves" in addition to the "smart moves," but the stupid moves, being failures, disappeared from view. All we see is the unbroken string of triumphs.[6] When we set ourselves the task of explaining why *those* were the triumphs, we uncover the reasons for things—the reasons already "acknowledged" by the relative success of organisms endowed with those things.

The original reasons, and the original responses that "tracked" them, were not ours, or our mammalian ancestors', but Nature's. Nature appreciated these reasons without representing them.[7] And the design process itself is the source of our own intentionality. We, the reason-representers, the self-representers, are a late and specialized product. What this representation of our reasons gives us is foresight: the real-time anticipatory power that Mother Nature wholly

6. This illusion has the same explanation as the illusion exploited by con artists in "the touting pyramid" (Dennett 1984d, pp. 92ff). Schull (forthcoming) argues that the process of natural selection need not always be *perfectly* stupid, brute force trial and error of all possibilities. Thanks to the Baldwin effect, for instance, species themselves can be said to pretest some of the possibilities in phenotypic space, permitting a more efficient exploration by the genome of the full space of the adaptive landscape. Just as creatures who can "try out options in their heads" before committing themselves to action are smarter than those merely Skinnerian creatures that can only learn by real-world trial and error (Dennett 1974a), so species that "try out options in their phenotypic plasticity" can—without any Lamarckian magic—give Mother Nature a helping hand in their own redesign.

7. Pursuing Schull's (forthcoming) extension of the application of the intentional stance to species, we can see that in one sense there is representation in the process of natural selection after all, in the history of variable proliferation of phenotypic "expressions" of genotypic ideas. For instance, we could say of a particular species that various of its subpopulations had "evaluated" particular design options and returned to the species' gene pool with their verdicts, some of which were accepted by the species.

lacks. As a late and specialized product, a triumph of Mother Nature's high tech, our intentionality is highly derived, and in just the same way that the intentionality of our robots (and even our books and maps) is derived. A shopping list in the head has no more intrinsic intentionality than a shopping list on a piece of paper. What the items on the list mean (if anything) is fixed by the role they play in the larger scheme of purposes. We may call our own intentionality real, but we must recognize that it is derived from the intentionality of natural selection, which is just as real—but just less easily discerned because of the vast difference in time scale and size.

So if there is to be any original intentionality—original just in the sense of being derived from no other, ulterior source—the intentionality of natural selection deserves the honor. What is particularly satisfying about this is that we end the threatened regress of derivation with something of the right metaphysical sort: a *blind* and *unrepresenting* source of our own sightful and insightful powers of representation. As Millikan (unpublished, ms. p. 8) says, "The *root* purposing here must be unexpressed purposing."

This solves the regress problem only by raising what will still seem to be a problem to anyone who still believes in intrinsic, determinate intentionality. Since in the beginning was *not* the Word, there is no text which one might consult to resolve unsettled questions about function, and hence about meaning. But remember: the idea that a word—even a Word—*could* so wear its meaning on its sleeve that it could settle such a question is itself a dead end.

There is one more powerful illusion to scout. We think we have a good model of *determinate,* incontrovertible function because we have cases of conscious, deliberate design of which we know, in as much detail as you like, the history. We *know* the *raison d'être* of a pocket watch, or of a laying hen, because the people who designed (or redesigned) them have told us, in words we understand, exactly what they had in mind. It is important to recognize, however, that however incontrovertible these historical facts may be, their projections into the future have no guaranteed significance. Someone might set out with the most fervent, articulate and clear-sighted goal of making a pocket watch and succeed in making something that was either a terrible, useless pocket watch or a serendipitously superb paperweight. Which is it? One can always insist that a thing is, essentially, what its creator set out for it to be, and then when the historical facts leave scant doubt about that psychological fact, the identity of the

thing is beyond question. In literary criticism, such insistence is known, tendentiously but traditionally, as the Intentional Fallacy. It has long been argued in such circles that one does not *settle* any questions of the meaning of a text (or other artistic creation) by "asking the author." If one sets aside the author, the original creator, as a definitive and privileged guide to meaning, one can suppose that subsequent readers (users, selecters) are just as important signposts to "the" meaning of something, but of course they are just as fallible—if their endorsements are taken as predictors of *future* significance—and otherwise their endorsements are just more inert historical facts. So even the role of the Pepsi-Cola franchise holder in selecting the two-bitser *as* a q-balber is only one more event in the life history of the device in as much need of interpretation as any other—for this entrepreneur may be a fool. Curiously, then, we get *better* grounds for making reliable functional attributions (functional attributions that are likely to continue to be valuable aids to interpretation in the future) when we ignore "what people say" and read what function we can off the discernible prowesses of the objects in question, rather than off the history of design development.

We cannot begin to make sense of functional attributions until we abandon the idea that there has to be one, determinate, *right* answer to the question: What is it for? And if there is no deeper fact that could settle that question, there can be no deeper fact to settle its twin: What does it mean?[8]

Philosophers are not alone in their uneasiness with appeals to optimality of design and to what Mother Nature must have had in mind. The debate in biology between the adaptationists and their critics is a different front in the same edgy war (see chapter 7). The kinship of the issues comes out most clearly, perhaps, in Stephen Jay Gould's reflections on the panda's thumb. A central theme in evolutionary theory, from Darwin to the present (especially in the writings of François Jacob (1977) on the *bricolage* or "tinkering" of evolutionary design processes, and in those of Gould himself) is that Mother Nature is a satisficer, an opportunistic maker-do, not "an ideal engineer" (Gould 1980, p. 20). The panda's celebrated thumb "is not,

8. Quine's thesis of the indeterminacy of radical translation is thus of a piece with his attack on essentialism; if things had real, intrinsic essences, they could have real, intrinsic meanings. Philosophers have tended to find Quine's skepticism about ultimate meanings much less credible than his animadversions against ultimate essences, but that just shows the insidious grip of meaning rationalism on philosophers.

anatomically, a finger at all" (p. 22), but a sesamoid bone of the wrist, wrest from its earlier role and pressed into service (via some redesigning) *as* a thumb. "The sesamoid thumb wins no prize in an engineer's derby . . . But it does its job." (p. 24) That is to say, it does its job *excellently*—and that is how we can be so sure what its job is; it is obvious what this appendage is *for*. So is it just like the q-balber that began life as a two-bitser? Gould quotes Darwin himself:

> Although an organ may not have been originally formed for some special purpose, if it now serves for this end we are justified in saying that it is specially contrived for it. On the same principle, if a man were to make a machine for some special purpose, but were to use old wheels, springs,and pulleys, only slightly altered, the whole machine, with all its parts, might be said to be specially contrived for that purpose. Thus throughout nature almost every part of each living being has probably served, in a slightly modified condition, for diverse purposes, and has acted in the living machinery of many ancient and distinct specific forms.

"We may not be flattered," Gould goes on to say, "by the metaphor of refurbished wheels and pulleys, but consider how well we work." (p. 26) From this passage it would seem that Gould was an unproblematic supporter of the methodology of reading function off prowess—which is certainly what Darwin is endorsing. But in fact, Gould is a well-known critic of adaptationist thinking, who finds a "paradox" (p. 20) in this mixture of tinkering and teleology. There is no paradox; there is only the "functional indeterminacy" that Dretske and Fodor see and shun. Mother Nature doesn't commit herself explicitly and objectively to *any* functional attributions; all such attributions depend on the mind-set of the intentional stance, in which we assume optimality in order to interpret what we find. The panda's thumb was no more *really* a wrist bone than it is a thumb. We will not likely be discomfited, in our interpretation, if we consider it *as* a thumb, but that is the best we can say, here or anywhere.[9]

9. We can complete our tour of two-bitser examples in the literature by considering Sober's discussion (1984) of the vexing problem of whether to call the *very first* dorsal fins to appear on a Stegosaurus an adaptation *for cooling:*

> Suppose the animal had the trait because of a mutation, rather than by selection. Can we say that the trait was an adaptation *in the case of that single organism?* Here are some options: (1) apply the concept of adaptation to historically persisting populations, not single organisms; (2) allow that dorsal fins were an adaptation for the original organism because of what happened later; (3) deny that dorsal fins are adaptations for the initial organism but are adaptations when they occur in subsequent organisms. My inclination is to prefer choice 3. (p. 197)

See also his discussion of the functional significance of the skin-thickness of *Drosophila* moved to different environments (pp. 209–10), and his discussion (p. 306) of how one might figure out which properties are being selected *for* by Mother Nature (now in the guise of Dawkins's crew coach): "Was the coach selecting for combinations of rowers? Was he selecting for particular rowers? We need not psychoanalyze the coach to find out." Not psychoanalysis, but at least the adoption of the intentional stance will help us do the reverse engineering we need to do to get any answers to this question.

After all these years we are still just coming to terms with this unsettling implication of Darwin's destruction of the Argument from Design: there is no ultimate User's Manual in which the *real* functions, and *real* meanings, of biological artifacts are officially represented. There is no more bedrock for what we might call original functionality than there is for its cognitivistic scion, original intentionality. You can't have realism about meanings without realism about functions. As Gould notes, "we may not be flattered"—especially when we apply the moral to our sense of our own authority about meanings—but we have no other reason to disbelieve it.