Back from the Drawing Board
Daniel Dennett

Reading the essays that make up this volume has shown me a great deal, both about the substantive issues I have dealt with and about how to do philosophy. On the former front, they show that I have missed some points and overstated others, and sometimes just been unable to penetrate the fog. On the latter front, they show how hard it is to write philosophy that *works* - and this is the point that stands out for me as I reflect on these rich and varied essays. Philosophical books and articles routinely fail to achieve their manifest goal of persuading their intended audiences of their main points. Does this make philosophy any worse off than other writing endeavors? Most published novels are failures of one sort or another, and the press has reported a recent study (whose methodology I wonder about) that concludes that the *median* Humber of readers of any paper published in a psychology journal is zero. But it seems to me that philosophy displays a particularly feckless record, with such a huge gap between authorial pretense and effect achieved that it is perhaps more comic than pathetic. In one weak moment I found myself thinking that perhaps some of our French colleagues have the right idea: deliberate obscurantism and the striking of stylized poses - since the goal of persuading by clear, precise analysis and argument is so manifestly beyond us. But that world-weariness passed, I'm happy to *Say*, and my cock-eyed American optimism returned. My ambition continues to be to change people's minds, and not just to win people over to my way of doing philosophy, as Bo Dahlbom suggests. But I admit that it is harder than I had thought. It's hard enough to get a good idea, but sometimes it's even harder, apparently, to get others to see what the idea is, and why it's good.

I view philosophical writing in an engineering spirit: I design and build devices - arguments, intuition pumps, metaphor machines - that are supposed to achieve certain effects. When they work, I'm exultant and when they don't, I'm surprised and annoyed. These essays uncover a variety of failures - mostly mine - and now that I have worked through my initial annoyance, I am ready, like any good engineer, to face the question: why didn't this thing work? What over-optimistic assumptions led me to underdesign the device? Is it worth repairing, or should I junk it and try another tack?

1 Science as a Virus - and Other Metaphors

Let's begin with one of my most striking failures, the twice-quoted passage on metaphor that closes *Consciousness Explained*. Dick Rorty loves it and Bo Dahlbom is, well, almost shocked by it. ("Quine would *never* have said that.") Neither understands it in the way I intended it. I was actually trying for quite a modest little point in that passage. I was not trying to say It's All Just Metaphor or anything remotely like that. I was just pointing out that however precise, literal, and non-metaphorical science may be in *some* of its accounts, it is never free of metaphors; some are good, some bad, and all are potent. I was proposing to replace some of the bad ones with some better ones.

I was not trying to hit a home run. Philosophers are supposed to try to hit home runs, and I guess sometimes that's why I am misunderstood. I seldom swing for the fences; I try to scratch out my runs by a more modest collection of bunts, grounders, and aggressive base running. I'm not noted for modesty - "hubris" and "arrogance" are words that appear all too often in reviews of my books - but in one regard I am more modest than philosophers generally are. I am shy about drawing ultimate conclusions about Reality, Truth, Meaning, Time, Causation, and the other grand topics of metaphysics and epistemology. Those questions are too hard for me to approach with any confidence. I take myself rather to be just working out some of the more surprising implications of the standard scientific picture, roughly the one I was given in school, and I don't even include all of that standard picture. As Dahlbom notes, physics is "none of my business," because I don't think I understand the niceties of contemporary physics well enough to be comfortable holding forth about its ontological presuppositions, and so forth.

Attitudes towards science are at the heart of several of the disagreements expressed in these essays, and perhaps the place to begin is with the contrast between Rorty and Richard Dawkins. The last section of Dawkins' idea-packed essay asks if science is a virus, and answers:

> No. Not unless all computer programs are viruses. Good, useful progams spread because people evaluate them, recommend them and pass them on. Computer viruses spread solely because they embody the coded instructions: "Spread me." Scientific ideas, like all memes, are subject to a kind of natural selection, and this might look superficially virus-like. But the selective forces that scrutinize scientific ideas are not arbitrary or capricious. They are exacting, well-honed rules, and ... they favor all the virtues laid out in textbooks of standard methodology: testability, evidential support, precision, ... and so on (p. 26).

When you examine the reasons for the spread of scientific memes, Dawkins assures us, "you find they are good ones." This, the standard, official position of science, is undeniable in its own terms, but question-begging to the mullah and the nun - and to Rorty, who would quite appropriately ask Dawkins: "Where is your demonstration that these 'virtues' are *good* virtues? You note that people evaluate these memes and pass them on - but if Dennett is right, people (persons with fully fledged selves) are themselves in large measure the creation of memes – something implied by the passage from Dennett you use as your epigraph. How clever of some memes to team together to create meme-evaluators that favor *them!* Where, then, is the Archimedean point from which you can deliver your benediction on science?”[1]

There is none. About this, I agree wholeheartedly with Rorty. But that does not mean (nor should Rorty be held to imply) that we may not judge the virtue of memes. *We* certainly may. And who are we? The people created by the memes of Western rationalism. It does mean, as Dawkins would insist, that certain memes go together well in families. The family of memes that compose Western rationalism (including natural science) is incompatible with the memes of all but the most pastel versions of religious faith. This is commonly denied, but Dawkins has the courage to insist upon it, and I stand beside him. It is seldom pointed out that the homilies of religious tolerance are tacitly but firmly limited: we are under no moral obligation to tolerate faiths that permit slavery or infanticide or that advocate the killing of the unfaithful, for instance. Such faiths are out of bounds. Out of whose bounds? Out of the bounds of Western rationalism that are presupposed, I am sure, by every author in this volume. But Rorty wants to move beyond such parochial platforms of judgment, and urges me to follow. I won't, not because there isn't good work for a philosopher in that rarefied atmosphere, but because there is still so much good philosophical work to be done closer to the ground.

Like most cognitive scientists, I'm prepared to take my chances with conservative, standard scientific ontology and epistemology.[2] My project, then, is to demonstrate how a standard, normal respect for science - *Time Magazine* standard, nothing more doctrinaire - leads inexorably to my views about consciousness, intentionality, free will, and so forth. I view science not as an unquestionable foundation, but simply as a natural ally of my philosophical claims that most philosophers and scientists would he reluctant to oppose. My "scientism" comes in the form of a package deal: you think you can have your everyday science and reject my "behaviorism" as too radical? Think again.

Now although I'm not swinging for the fences, I might happen - almost by accident - to hit a home run once in a while. John Haugeland sees me as contributing to ontology, the most Olympian feat in metaphysics, and Rorty and Dahlbom find similar implications to discuss. Maybe Rorty is right that I *ought* to be more ambitious and radical in my conclusions. More on that later, but in the meantime, there is plenty of good work to be done on the basepaths - and plenty of room for instructive failure.

2 A Case from Cognitive Neuroscience: Churchland and Ramachandran

A clear instance of this is found in Pat Churchland and "Rama" Ramachandran's essay (chapter 2) on the metaphor of "filling in", which is ubiquitous in cognitive science, but which I have attempted to condemn. Their essay shows that I'm certainly wrong about one thing: about the effects my own metaphor-machines would have on them, for I have misled them. They don't understand what I was trying to assert (and deny), and if *they* don't understand, what chance do I have of getting through to the uninitiated?

The issues here are not grand metaphysical issues, but rather nagging confusions about how to think about vision. We (scientists and philosophers) have to use metaphors to think about something as complex and puzzling as vision, but the metaphors can also create difficulties for us. The way to overcome these difficulties is not from on high, with sweeping philosophical theses, but in the trenches, looking at the details of how we describe the phenomena and what assumptions we make about them in the process.

They propose to sweep the decks of "needless metaphysical tut-tutting" by claiming that their talk of filling in is only a convenient shorthand; speaking this way does not commit them to any silly view that holds that when someone sees an apple "there might be a little (literal) apple or a (literal) picture of an apple in someone's head which is the thing that is filled in." Good. Then to what does their convenient shorthand refer? "Merely to some property of the brain's visual representation such that the perceiver sees a non-gappy apple." But what does *that* mean? This is the whole issue.

They begin by confronting a choice, an either/or with two alternatives. What is going on, they ask, when one's blind spot is seen as filled in? "Is it analogous to acquiring the non-visual representation (belief) that Dowser, the family dog, is under the bed. . . Or is it more akin to regular visual perception of the whole Bowser in one's *peripheral but non-blind field?* That is, is the representation itself a visual representation, involving visual experiences?" (p. 30). The contrast is not as clear as they may think; there seem to me to be intermediate alternatives. The difference they are alluding to is exemplified, one gathers, by the difference between coming to believe that there is a dog under the table by *deducing* this from various non-visual clues (including, perhaps, a trusted friend whispering in your ear. "There's a dog under the table!") and coming to believe there is a dog under the table by *seeing* it in one's peripheral vision. The latter case, but not the former, involves "visual experiences." And if you just see the dog's tail out of the corner of your eye and infer the rest, this is presumably a mixed case, partly involving visual experience, partly involving visually unclothed inference. What, then, about watching a dog walk behind a picket fence? There is no deliberate, conscious deduction of the hidden bits, certainly, but in some sense they are inferred - *and* your brain accomplishes this (I reckon) without having to draw in, or paint in, or in any other way fill in the hidden bits. You don't "just think" there's a whole dog; you *see* that it's a whole "non-gappy" dog walking behind the fence; this is certainly a "visual experience" even though its completion (so that you see what's in front of you *as* a whole connected dog) involves jumping to a conclusion about what is missing.

At any rate, this is how I would put it, but this courts misunderstanding. Churchland and Ramachandran take the contrast to be between peripheral visual perception and entirely non-visual perception (or just belief), but I am just as insistent that there is no "filling in" in normal peripheral vision as in the blind spot. For instance, there is nothing "non-visual" about your discovery, on entering a room, that it is covered wall-to-wall with identical photos of Marilyn Monroe, but that discovery *must* be way out ahead of the information your eyes have taken in, since your peripheral visual system is simply unable to distinguish the peripheral Marilyns from Marilyn-shaped blobs. I count "jumping to a conclusion" (e.g., jumping to the conclusion that the room is papered with identical Marilyns) as the brain's "doing something positive" but not as "filling in" because, although the brain does indeed add information as it draws the conclusion, it doesn't then go on to draw the Marilyns - nor is its conclusion based on any extrapolative drawing it has already done. Or so I claim.

This is an empirical issue, and there are ways that I could be shown to be wrong. The cases Churchland and Ramachandran describe illuminate the issues, and in some cases non-trivial clarifications to my position, but for the most part they are either neutral or support my minimalist view, not their alternative. That this should not yet be obvious to Churchland and Ramachandran shows how far from clarity my framing of the issue has been to date.

Consider the case of the field of bagels (their Figure 2.8), in which one bagel is carefully positioned so that its "hole" falls within the subject's blind spot. If they had asked me in advance of the experiment whether the subject would see "just more bagels" or rather a solid disk, I would not have known what to say; I would have recognized an ambiguity in my claim that the brain assumes "more of the same," for it all depends on whether the brain draws this conclusion locally or globally. They assume that my view would have to predict that the bagel would be seen *as* a bagel (globally more of the same), but this does not follow. One can see the alternative by looking at a succession of different cases. Suppose your visual field consisted of solid red; then drawing the conclusion "more of the same" would lead to the same result whether the reasoning was based on global or local conditions. Suppose next that there is a single big red disk off to the side in your visual field. If the brain simply concludes that the property it finds *in the general vicinity* of the blind spot - red - is to be extrapolated, this is a case of concluding "more of the same," and does not also have to involve filling in. Now shrink the disk to the size of a bagel- and, if you like, punch a hole in it, covered by the blind spot. If the brain bases its conclusions on what is in the *immediate vicinity* of the blind spot, then "more of the same" gives just more red, so the brain will indeed conclude it has a solid red disk, which might then "pop out" just as in Ramachandran's experiment. So what the experiment shows is that at least in this instance, the brain bases its extrapolation on local conditions, not global conditions. This is a valuable result; in retrospect we may say it is not surprising, but retrospect doesn't count. I would have said it could go either way before I learned of the experiment, which does indeed show that the brain's conclusion (in this instance) is both locally based and "early," early enough in visual processing to provide the material for "pop out." (One might have predicted that the bagel would be seen as a disk but would *not* pop out, on the supposition that the conclusion *red disk* would be reached too late to provide an odd-one-out premise for pop-out.)

The experiment in which subjects complete the bar through the blind spot, in spite of the fact that the top half is red and the bottom is green, raises a different issue. "Subjects still see the bar as complete, with extensions of both the red and green bar, but they do not see a border where the red and green meet, and hence they cannot say just where one color begins and the other leaves off." Now this way of putting it blurs the distinction I am trying to make. Is it just that "they cannot say" where one color begins and the other leaves off, or is it that the brain itself never "says" where one color begins and the other leaves off? If there is *any* sort of filling in worthy of the name, then each sub-area of the bar-as-represented must be filled in either red or green (or "reddish green" as in the Crane and Piantanida experiment!). Or I suppose areas could flicker back and forth between red and green, but one way or another, filling in requires explicit representation of the color at each "pixel" within the outline of the bar - that is what I mean by filling in (sec *Consciousness Explained* p. 349). But if there isn't filling in, if the brain just concludes that it is a single solid bar with a red top and a green bottom and *does not go into the matter* of where and how the color changes, then there would be no fact of the matter about where "the boundary" was, just as there is no fact of the matter about whether Falstaff had a sister. The brain might even in some sense "recognize" that there had to be a boundary between the two colors (Rule violation! Two color labels on a single region!), but just neglect to resolve that violation and settle where that boundary might be. It is not clear, from Churchland and Ramachandran's formulation, whether they suppose that there must be a boundary in the brain's representation, a boundary that is simply inaccessible to the subjects. I am claiming that while there *might* be - it is an empirical question - there *need not* be. From my reading of other work of Ramachandran's, I expect he agrees, and perhaps Churchland does as well.

The conclusion Churchland and Ramachandran draw from the blind-spot experiments is that "in so far as there is nothing in the visual stimulus corresponding to the filled in perception, it is reasonable to infer, in contrast to Dennett, that the brain is 'providing' something, not merely 'ignoring something.'" But their claim does not in fact contrast with mine - or perhaps I should say, it does not contrast with what I intended to convey. The brain certainly provides content regarding the blind-spot region, but it may still be accomplishing this by "ignoring." To ignore an area is to operate without requiring confirmation or disconfirmation from the area - going along as if that bit of space just didn't exist. Jumping to the conclusion that it's more of the same is "adding something" - it is to be distinguished, after all, from not jumping to that conclusion at all - but it is not filling in. The difference I am after is the difference between jumping to a conclusion and stepping to a conclusion by making some bogus steps on which to rest the conclusion (e.g., paint in the region, and then use that painted-in region as one's "evidence" for the conclusion subsequently drawn). The way to test my hypothesis that the brain does not bother filling in the "evidence" for its conclusion is to see if there are effects that depend on the brain's having represented the *step*, rather than just the *conclusion*.

How could this be shown? We might look for inspiration to the famous random-dot stereogram experiments by Bela Julesz. Before Julesz demonstrated that stereo vision could be achieved by simply displacing a region of random dots in the view presented to one eye (creating binocular disparity, interpreted as depth), a minimalist creed prevailed: information about the individual dots on the retinas was not preserved up to the optic chiasm (where the signal streams from both eyes merge for the first time). The brain, it was thought, had already thrown away such pixel-by-pixel information, and replaced it with generalizations, along the lines of "more random dots over here." But since depth perception was demonstrably possible in Julesz's conditions, the pixel-by-pixel information *had* to be preserved this "late" in the visual stream, since if the brain were merely noting that there were "more random dots" there would be no individual dot-representations to match up in order to identify the displaced area. Julesz in effect demonstrated that *each dot* was represented by the information in each stream.

A similarly striking demonstration might prove that the brain actually does fill in pixel-by-pixel values in the region representing the blind spot (or other regions, such as artificial scotomas). The detail would not just *seem* to be there; it would *have* to be there to explain some effect. (Note that Churchland and Ramachandran's essay offers fine examples of the importance of adopting what I call the method of *heterophenomenology (Consciousness Explained*, ch. 4). Consider the use they make of subject report; it is not enough to predict phenomenology; you have to show whether the subjects are right about it; you mustn't just assume they are.) I have not yet been able to dream up a plausible experiment that could show this, but Ramachandran will be able to, if anyone can. There are some positive but inconclusive hints in the data he has already presented: the long latency of some filling-in effects, and even more important, the *spatial spread,* such as "the red color bleeding into" the scotoma. This is telling (if not quite conclusive), because you can't assign a label to a region *gradually,* and it is hard to think of a reason why the brain would bind the "gradually" label to a process that wasn't gradual. (It *seems* to be gradual, but maybe it *only* seems to be gradual.)

The Gattass effect and the Gilbert and Wiesel effect, which apparently show that there is much more dynamic migration of receptive fields than heretofore suspected, do indeed add a new set of options to the standard ways of thinking about how the visual system might or must work. As Churchland and Ramachandran say, the options are not exhausted by my "bit map or color-by-numbers" alternatives, but those were not meant to be exhaustive: they were merely meant to be alternative ways of thinking of the *traditional* issue of "filling in." These recent neuroanatomical discoveries are important because they show another possibility: treating input from one area as if it came from another. But Churchland and Ramachandran fail to notice that there are (at least) two very different ways the brain could do this, which we might crudely label "duplication" and "pinching off."

In duplication, the input from one small region is actually *doubled* at a higher level, to represent both its home region and an adopted region. "Pinching off" is a more radical hypothesis: one's visual field can be represented as a space that has no unique mapping onto any portion of the ordinary three-dimensional space in front of your eyes; gaps in one's coverage of that space can be simply "pinched off" in one's visual representation, giving the sense of a seamless, gapless space that simply *leaves out* certain regions of the external world. If I may risk a metaphor: think of printing a map of the USA on a soft rubber sheet, and then, from behind, simply pinching Tennessee out of sight, bringing Kentucky's southern border right down to the northern borders of Mississippi and Alabama, and then smoothing out this infolding, so that a seamless plane resulted. Is that "filling in?" It might be what the brain sometimes does. Ramachandran's most recent researches, conducted after the essay in this volume was completed, show that duplication can occur in other sense modalities, such as phantom limb, so it is possible, but not yet proven, that such duplication occurs in the case of the blind spot.

Each of the other experiments Churchland and Ramachandran describe deserves an equally detailed discussion from me, but since I think I have now expressed all the main themes those detailed discussions would appeal to, instead of "filling in" those details, I will encourage the reader to extrapolate "more of the same" and see if that doesn't provide an adequate response. Churchland and Ramachandran chide me for thinking like a computer engineer and prejudging how the brain might or must deal with its tasks. I plead *nolo contendere.* It's a good way to think - risky, but a fine way of generating hypotheses. I might sometimes forget to admit, in the thrill of the chase, that these hunches need confirmation (and hence court disconfirmation), and, of course, I must take my lumps when I'm proven wrong. But although the results they describe require me to clarify and disambiguate my hypotheses, they don't *yet* require me to retract them.

The pure philosophers may want to come up for air at this point. Churchland and Ramachandran provide a closing diagnosis of where I went wrong (by their lights) which imputes to me a variety of *behaviorism* which they think has blinded me to the importance of the neuroanatomical details. They contrast my view with "scientific realism," according to which "it is reasonable to consider sensory experiences to be real states of the brain, states whose neurobiological properties will be discovered as cognitive neuroscience proceeds." But I agree wholeheartedly with this, and am, indeed, as scientific a realist as one could find. While I am proud to trace my lineage to Quine and Ryle, I must protest that this identification of my position breeds more misapprehension than enlightenment.

3 Labels: Am I a Behaviorist? An Ontologist?

Am I a behaviorist? Tom Nagel and John Searle have always wanted to insist on it, and now Churchland concurs, and so does Dahlbom, who very accurately describes my position -but imposes on it a label I abhor. I agree that there is historical justification for the label, spelled out by Dahlbom, but most people don’t appreciate the subtleties of that justification, and hence are drastically misled by the term's connotations.

I once made the mistake of acquiescing, in a cooperative spirit, when Ned Block suggested that I should let myself be called an instrumentalist. Never again. While there are perfectly reasonable doctrines that in fact every right-thinking scientist holds that might reasonably he called instrumentalism (instrumentalism with regard to centers of gravity, instrumentalism with regard to parallelograms of force, etc.), tradition has it that instrumentalism is the all-too-radical dogma that treats electrons (and for that matter, whales and mountains) as in the same ontological boat as the Equator. So when I let myself be counted as an instrumentalist, I then found I had to work very hard to undo the damage. People quite naturally reasoned that if I was a self-professed instrumentalist, and if some dime-store argument refuted instrumentalism, the same argument must refute me. Since people are inclined to reason this way, the tug-of-war about labels (which "ism" do you pledge allegiance to?) is regrettably important.

We all go in for this sort of reasoning. If I learn that somebody is an idealist, or a dualist, my initial working assumption is going to be that this person holds a forlorn view, since the "refutations" of idealism and dualism are well known. So if I held a view that could be seen, in a certain light, to be a *sort* of dualism, I'd be extremely reluctant to "admit it," since the debates that ensued would so boringly gravitate towards defending my view against all the old arguments. The standard arguments against both Skinnerian and Rylean behaviorism do not touch my view; indeed, I am the author of some of those arguments ("Skinner Skinned," "A Cure for the Common Code"). My anti-behaviorist credentials are impeccable.

But people like a memorable label for a view, or at least a slogan, so since I reject the label, I'll provide a slogan: "Once you've explained everything that happens, you've explained everything." Now is that behaviorism? No. If it were, then all physiologists, meteorologists, geologists, chemists, and physicists would be behaviorists, too, for they take it for granted that once they have explained all that happens regarding their phenomena, the job is finished. This view could with more justice be called phenomenology! The original use of the term "phenomenology" was to mark the cataloguing of everything that happened regarding some phenomenon, such as a disease, or a type of weather, or some other salient source of puzzlement in nature, as a useful preamble to attempting to explain the catalogued phenomena. First you accurately describe the phenomena, as they appear under all conditions of observation, and then, phenomenology finished, you - or someone else - can try to explain it all.

So my heterophenomenology is nothing more nor less than old-fashioned phenomenology applied to people (primarily) instead of tuberculosis or hurricanes: it provides a theory-neutral, objective catalogue of what happens - the phenomena to be explained. It does assume that all these phenomena can be observed, directly or indirectly, by anyone who wants to observe them and has the right equipment. It does not restrict itself to casual, external observation; brain scans and more invasive techniques are within its purview, since *everything that happens in the brain* is included in its catalogue of what happens. What alternative view is there? There is only one that I can see: the view that there are subjective phenomena beyond the reach of *any* heterophenomenology. Nagel and Searle embrace this curious doctrine. As Rorty notes: "Nagel and Searle see clearly that if they accept the maxim, 'To explain all the relational properties something has - all its causes and all its effects - is to explain the thing itself;' then they will lose the argument" (p. 185). They will lose and science will win.

Do you know what a zagnet is? It is something that behaves exactly like a magnet, is chemically and physically indistinguishable from a magnet, but is not really a magnet! (Magnets have a hidden essence, I guess, that zagnets lack.) Do you know what a zombie is? A zombie is somebody (or better, something) that behaves exactly like a normal conscious human being, and is neuroscientifically indistinguishable from a human being, but is not conscious. I don't know anyone who thinks zagnets are even "possible in principle," but Nagel and Searle think zombies are. Indeed, you have to hold out for the possibility of zombies if you deny my slogan. So if my position is behaviorism, its only alternative is *zombism.*

"Zagnets make no sense because magnets arc just *things* - they have no inner life; consciousness is different!" Well, that's a tradition in need of reconsideration. I disagree strongly with Rorty when he says "Dennett's suggestion that he has found neutral ground on which to argue with Nagel is wrong. By countenancing, or refusing to countenance, such knowledge, Nagel and Dennett beg all the questions against each other" (p. 188). I think this fails to do justice to one feature of my heterophenomenological strategy: I let Nagel have everything he wants about his own intimate relation to his phenomenology *except* that he has some sort of papal infallibility about it; he can have all the ineffability he wants; what he can't have (without an argument) is *in principle* ineffability. It would certainly not be neutral for me to cede him either infallibility or ineffability *in principle.* In objecting to the very idea of an objective standpoint from which to gather and assess phenomenological evidence, Nagel is objecting to neutrality itself. My method does grant Nagel neutral ground, but he wants more. He won't get it from me.

Are there any good reasons for taking zombies more seriously than zagnets? Until that challenge is met, I submit that my so-called behaviorism is nothing but the standard scientific realism to which Churchland and Ramachandran pledge their own allegiance; neither of them would have any truck with phenomenological differences that were beyond the scrutiny of any possible extension of neuroscience. That makes them the same kind of "behaviorist" that I am - which is to say, not a behaviorist at all!

Labels do seem to have given me a lot of trouble. One of the most frustrating is "realism." Rorty urges me to stop taking the controversy over "realism" seriously. 1 wish 1 could, but Haugeland, Fodor and Lepore, and others tug me in the opposite direction. My penchant, so tellingly described by Dahlbom in the introduction, for deflecting objections with storytelling and analogy-drawing instead of responding with clearly marshalled premises and conclusions, sometimes backfires, as I learned in a recent discussion with that metaphysician *par excellence*, Jaegwon Kim. He had complained to me roughly as follows:

> you just give examples and leave it to us to draw the moral; we want you to say what conclusions, what general principles, you think they establish. Instead of answering by stating and defending an ontological position, you just give another example. The examples are very nice, but we want the theory that accounts for them.

My reply: that is not quite the point, of my examples. We've been at crosspurposes, for 1 have not been sufficiently forthright in saying what I take the force of my examples to be. My point is to respond to the challenge from the metaphysicians or ontologists by saying, in effect, Look, when you folks get clear about such non-mysterious cases as smiles, opportunities, centers of gravity and voices, then (and only then) will I feel the obligation to answer your questions. Until that time, the ontological enterprise is nothing I feel obliged to engage in. Are centers of gravity abstract or concrete? Should one be an eliminative materialist about smiles, and then commit oneself to *properties*, the instantiations of which are *events?* Or should one be a realist about them? Are these options ever really serious? I am sure any ontologist worth his salt can provide a rigorous, clear, systematic account of all these cases. (I have tried it out myself, in private, and have some ideas about what might work.) The trouble is that different ontologists come up with different systems, and they are unable to reach consensus. I really am prepared to take instruction on these issues, but not until the instruction is in unison or close to it. I wouldn't want to trot out *my* ontology and then find I had to spend the rest of my life defending or revising *it*, instead of getting on with what are to me the genuinely puzzling issues - like the nature of consciousness, or selves, or free will. The ontological status of fictional characters, haircuts, holes, and the North Pole may be deep and fascinating problems in themselves to some philosophers, but not to me; they are interesting playthings to come back to in one's spare time. *That* is the game I am opting out of, since there seem to me to be better things to do first. When and if professional ontologists agree on the ontological status of all my puzzle examples, my bluff will be well and truly called; I will feel a genuine obligation to make things clear to them in their terms, for they will have figured out something fundamental.

But John Haugeland insists that I am doing ontology whether I like it or not, and I ought to make sure I do it right. One thing is clear; it was at least a tactical mistake for me to describe my alternative to Realism and Eliminativism and Instrumentalism as "mild realism." Both Rorty and Haugeland jump on me for that, and I guess they are right. (Certainly Rorty is right that my description of his view in terms of "irrealism" does it an injustice.) But must I attempt to put my ontological house in order before proceeding further? I don't mean to suggest that this is just a matter of taste; I may be simply wrong about the value of getting ontology straight as a first step. I think, in fact, that ontologists differ in their creeds about this. Some, I gather, really don't care whether ontology is essential to scientific progress; they just find the puzzles intrinsically so interesting that they are prepared to devote their professional lives to trying to sort them out. I see nothing wrong with that attitude. I deplore the narrow pragmatism that demands immediate social utility for any intellectual exercise. Theoretical physicists and cosmologists, for instance, may have more prestige than ontologists, but not because there is any more social utility in the satisfaction of *their* pure curiosity. Anyone who thinks it is ludicrous to pay someone good money to work out the ontology of dances (or numbers or opportunities) probably thinks the same about working out the identity of Homer or what happened in the first millionth of a second after the Big Bang.

Of course, some cosmologists would insist that their researches do - or might well - have social utility in the long run, and in a similar spirit other ontologists seem to think what they are saying is directly relevant to, and important to, the scientists working on empirical theories, for instance, theories of the mind. But I have never seen a persuasive case made for this view, and so far as I can see, ontology has always in fact lagged behind science, whether or not this is a good thing. I have yet to see a case where *first* the metaphysicians got a clear account of the ontology and this *then* enabled the scientists to solve some problem by rendering theory in its terms.

This doesn't mean that I think science is conducted in ontology-neutral terms, or that the ontologies scientists tacitly adopt don't influence (even cripple) their scientific enterprises. Quite the contrary; I think ontological confusions are at the heart of the lack of progress in cognitive science. But I don't think the way to overcome the problem is by stopping the science until the ontology is clear. Here is where we really are in Neurath's boat, and must rebuild it while we keep it sailing.

How, then, do we rebuild it, if not by first developing a "systematic" ontology? By noting pitfalls, looking at analogies, keeping examples close to our attention, etc. This is what I was attempting to do in "Real Patterns," but Haugeland finds my attempt hampered by two sorts of vacillation. In the first instance, I vacillate between talking about the reality of a pattern, and the reality of the elements of the pattern. That's a fair cop. When I asked "When are the elements of a pattern real and not merely apparent?" I meant to be asking about the reality of what we might better call the pattern's *features* (which only exist *as* features-of-a-pattern), since the reality of the ultimate elements of which these features are somehow composed was not at issue. Haugeland thinks this slip of mine interacts in a significant way with the second sort of vacillation he spies, but I am not persuaded. The second vacillation is between two different definitions of pattern: one in which patterns are "recognizabilia" as he puts it, and one that aspires to objectivity (independence from the biases and limitations of any particular recognizer). He is surprised that I don't note the tension between these; in fact, I thought I had resolved the tension and unified them via the mathematical definition of randomness, but I grant that I should have spelled it out. A mathematically random sequence is not a pattern, and any patterns "discerned" in it are strictly illusory, not real; and any sequence that is not mathematically random has a pattern that is recognizable-in-principle by some observer or other, up to and including what we might call the maximal observer: the whole universe considered as a potential observer. Doesn't this keep the tie to recognition while achieving objectivity?

Of course Haugeland wants to treat recognition in its richest garb, as something more than mere discrimination and differential reaction, which is all I wanted to mean by the term. I should perhaps have chosen a more modest or technical term, such as pattern-*transduction*, instead of relying on the standard idiom as it appears in the literature of AI and cognitive science. But then I mightn't have provoked Haugeland's interesting analysis of the foundational role of (rich) recognition in ontology, so it is a fortunate error. One particularly important point he makes is that recognition does not always depend on a process of analysis of elements, but I don't think he draws the right conclusion from it. For this is true even in those cases where the difference between correct and incorrect recognition *can be defined in terms of* such elements. The difference between the written words "boat" and "coat" is nothing other than the occurrence of "b" at the head of one and "c" at the head of the other, but it does not follow that particular acts of visual recognition of either word consist of independent acts of "b" -recognition and "c" -recognition and so forth. Often it is only the wider context that permits a perceiver to recognize a dimly or distantly seen word as "boat" *which then enables* the perception of the leftmost element *as* a "b," something otherwise quite beyond the perceptual capacity of the recognizer. It is even possible for there to be a "boat" -recognizer who hasn't a clue about – cannot recognize - the individual letters that are the definitive elements of "boat."

Haugeland thinks there is a difficulty for my proposed merger of the two senses of pattern because "if. . . an attempt were made to merge the two notions of pattern, such that recognizable patterns must at the same time *be* arrangements of prior elements, then, arguably, their recognizability would have to be *via* prior recognition of those elements ... " (p. 59). Arguably, but mistakenly arguably! I think I can accept his observations about the non-algorithmic nature of recognition, while clinging to the conservative default ontology: patterns are patterns of prior elements, even if you don't know what those elements are (yet). The reality of the pattern *features* depends on their being in principle recognizable as features by some pattern recognizer, but this recognizability does not in turn depend on some base recognizability of the atoms, as it were, of which those features are composed.

So I'm not convinced. But if I am wrong and Haugeland is right, then this would be a case in which something of a revolution in ontology came from within, boiling up out of the scientific details, not from on high. And, in the meantime, when people challenge me to announce my allegiance - to eliminative materialism or property dualism or realism or emergentism or even (God forbid) epiphenomenalism, or whatever - I will firmly resist the challenge. Of course, this means that I pass the buck to them; I say, in effect, *you tell me* what category my view falls in. Since I do this, I can't complain *that* they categorize me (I invited them to) but I can complain when they do it wrong. My view is *itself* one of my puzzle examples; I challenge *them* to come up with an ontological interpretation of it that does justice to its subtleties, just as I challenge them to do the same for centers of gravity or voices.

That, then, is the strategy I have adopted, but never before been so forthright about. My reticence has cost me some misconstruals, such as those of Jerry Fodor and Ernie Lepore, who work hard to turn me into something I am not. (Fodor believes that doing philosophy is a matter of composing Arguments from Principles, and when he encounters something that purports to be philosophy, he does his best to shoehorn it into that shape.)

Consider the thrashing they give my discussion of adopting the intentional stance towards "Mother Nature" - the process of natural selection. They find it "very puzzling." Why? Because it doesn't fit the mold. They try to turn it into an argument, which they can then rebut by parity of reasoning with their story about Father Erosion and the Tree Fairy. They end with the rhetorical question: "if the story about Father Erosion doesn't legitimize interpretivism about functions in ecology, why, exactly, does the story about Mother Nature legitimize interpretivism about functions in biology?"

This is interesting. Fodor and Lepore think my remarks about Mother Nature's reasons were intended as an *argument* to "legitimize interpretivism" when in fact they were intended to *explain* interpretivism, to make it comprehensible and palatable, not to prove it. How then do I "legitimize interpretivism about functions in biology?" By examining the assumptions and constraints encountered in actual attempts to impute functions in biology. What one discovers is that assertions about function are in the end answers to "why" questions. Such "why" questions can only be satisfied by giving reasons. Whose reasons? Well, nobody's reasons: free-floating rationales, I have called them, but call them what you like, they are ubiquitous and (almost) uncontroversial in biology. My reasons for interpretivism in biology are thus not derived from an argument from general principles. Rather, they are simply to be observed in the details: look and see - the function-ascribers always justify their claims by engaging in interpretation, invoking reasons. Is it acceptable to postulate such reasons that aren't any reasoner's reasons? Well, if it isn't, then we are really in trouble, since then *no* function attributions in biology are acceptable: we can't say the eagle's wings are for flying, or our eyes for seeing. There are indeed biologists who are strongly inclined to this puritanical conclusion, for they see that the only way to open the door to function attributions in biology is to countenance free-floating rationales.3

When casting about for materials for an Argument, Fodor begins, naturally enough, with the Principles that are pre-theoretically the most obvious (most obvious to his Granny, as he would say). Not surprisingly, Granny believes that beliefs are real, and really mean what they mean - "period!" as a Granny of my acquaintance would add. Formally rendered and capitalized, that's Intentional Realism.

Fodor and Lepore take themselves, then, to be mounting a Defense of Intentional Realism against its enemy, Interpretivism. Their essay proceeds systematically: first they divide the possible schools of Interpretivism into two camps: Projectivism and Normativism. Then they defeat Projectivism. Then they tackle Normativism, and find two candidate interpretations of my purported defense of it: an evolutionary argument or a transcendental argument. The former they find to yield the wrong conclusion, and the latter, they conclude, "begs the question against intentional laws."

We might call this method *philosophy drawn and quartered.* First you draw a line, dividing the world in half, and make everybody decide which half of the world they want to stand in; then you draw another line, drawing the half they've chosen in half, and demand that they choose one quarter or the other. And so forth. This imposition of either/or is supposed to guarantee that all philosophical positions have good hard edges. You force people off the fences they are sitting on, and make them choose sides for each imposed dichotomy. This does seem to make methodological sense from a certain perspective. The trouble with philosophy is often that it is vague, impressionistic and lacking in rigor, so this seems to be a good way to force precision onto the issues. Fish or cut bait, as the saying goes. Logic teaches us that if you take a well-formed proposition and its negation, exactly one of the pair is true and the other is false, so you'd think that drawing and quartering would be a way of homing in on the truth by a process of elimination.

The trouble is that false dichotomies are hard to spot. When you force people to jump to one side or the other of the fence before they have had a chance to reflect properly on the way the issue has been posed, they may end up not noticing the best alternative - some unimagined third or fourth alternative - because the "only options" are not really exhaustive contradictories, in spite of first appearances.

Fodor and Lepore provide an instructive case of this in their interpretation of the ontology of Interpretivism: "Interpretivism is, *inter alia*, the view that, strictly speaking, we don't really have beliefs and desires" (p. 74). Where do they get this? Well, logic teaches us that exactly one of the following two propositions is true: *beliefs are real* or *it is not the case that beliefs are real.* In other words, they conclude, "strictly speaking" ya gotta be a Realist or an Eliminativist - take your pick. They decide (not surprisingly, given that ultimatum) that Interpretivism falls in the Eliminativist quarter. Then they lower the boom: "Interpretivism is, *inter alia*, the view that, strictly speaking, we don't really have beliefs and desires. But, one supposes, what a creature *doesn't really have* can't help it much in its struggle for survival."

So much, then, for an evolutionary account of belief and desire if one is an Interpretivist. Well, consider an exactly parallel argument regarding centers of gravity. When it comes to centers of gravity, which are you - a Realist or an Eliminativist? Suppose you jump to the right, on the grounds that centers of gravity are "Interpretivist" or "Instrumentalist" posits, not real things like atoms or electrons: "Interpretivism is, *inter alia,* the view that, strictly speaking, nothing really has a center of gravity. But, one supposes, what a sailboat *doesn't really have* (e.g. a low center of gravity) can't help it much in its struggle against capsizing." Persuasive? I trust you feel the urge to back up a step or two and challenge the drawing and quartering that led to this embarrassing conclusion.

After concocting and demolishing a variety of views I don't hold, Fodor and Lepore eventually confront the view I do hold.

> Of course, many philosophers who think that charity constrains intentional ascription *a priori* doubt that there *are* intentional laws. We have nothing to say against their doubting this except that they are in need of an argument, and that, whatever this argument is, it mustn’t itself depend on assuming that charity is constitutive of intentional ascription ... In the present context, that assumption would be merely question-begging. (p. 79)

It seems they have cleverly maneuvered me into the deadliest trap in Burden Tennis (burden, burden, who has the burden of proof now?): they have got me where 1 need an argument (I have the burden), but where only a question-begging argument (in that context) is available. Darn, I lose.

But note that if one is not, at the moment, attempting to refute *X* but just explain what alternative *Y* is, it is quite acceptable to beg the question against *X*. For instance, any compact explanation of the role of DNA in biology would probably beg the question against vitalism. Life, after all, is short. I am accused of begging the question against intentional laws. Again I plead. *nolo contendere.* I had thought that the idea of intentional laws was so obviously mistaken that it was not worth arguing against.

5 Intentional Laws and Computational Psychology

Here is the issue that divides us. I have long claimed that the predictive value of belief ascriptions depends on our assuming the underlying (and pervasive) rationality of the believer. Fodor and Lepore do not believe this, and in the course of arguing against it, they say: "We'd get *some* predictive value out of belief ascription even if it only worked, say, 87 percent of the time that a creature that believes ($p \rightarrow q$ and p) believes q" (p. 77). This superficially plausible claim presupposes the independent discoverability of individual beliefs, whereas I hold that it is the assumption of rationality that makes belief-ascription "holistic": you can't attribute a single belief without presupposing the ascription of a host of others, rationally related to it. Let's look more closely at what would be involved in Fodor and Lepore's alternative "atomistic" Intentional Realism. They are supposing in this passage that we might *find* a case of somebody who believes both p and $p \rightarrow q$, and *then* determine independently whether or not he also believes q. Totting up a raft of such individual cases, we might discover that 87 percent of the time, people who believe p and $p \rightarrow q$ also believe q. But this assumes that determining, say, that a particular person didn't believe q wouldn't *ipso facto* undo or at least put into question one's earlier finding that the person believed p and $p \rightarrow q$. Sellars, Quine, Davidson, and I (among others) have always seen - or so we thought! - that individual belief attributions were not anchorable in this imagined way because one always encounters what we might call the Quinian circle: the very best evidence *there could be* that you had made a mistake in attributing to someone the beliefs that p and $p \rightarrow q$ was the evidence that you should also attribute the belief that *not-q.*

Imagine giving a person a huge true-false questionnaire to fill out, the 999-page F-SQUAB (Fodor Self-Questioning Attributor of Belief). Many of the sentences to be checked T or F are logically related to each other, and 10 and behold, 13 percent of the time, people who have checked T for pairs of sentences p and $p \rightarrow q$, check F for q. There would be controversy about whether these people *really believed* (or even understood) the sentences in question, given their anomalous performance on them, so we should review those sentences with the subjects, to get clarification. What could be better evidence that they had failed to understand (and hence believe) the sentences in question than their persistence in assigning T and F in the wrong way? Fodor would presumably answer: there *could* be better evidence - we could find those very sentences (or rather, their Mentalese translations written in the belief boxes inside their heads! (On this score, see the citation of Davidson by

Rorty, p. 192.) But identifying something in their heads as the belief box (and not just the memorized sentence box, for instance) is itself a hypothesis that raises the issue of rationality all over again. For one thing, we could never establish rules for translating Mentalese into a natural language without using precisely the assumption about rationality this move to the inside was supposed to avoid. Translating somebody's language of thought into, say, English, would itself be an exercise in Quinian radical translation.

Versions of this claim have been around - and unchallenged - for so long that I have been willing to presuppose it without further ado, but Fodor and Lepore demand an argument. If there were independently discoverable *intentional laws*, they suppose, we could use these as a base from which to get the leverage to break the Quinian circle. · There could be logically independent evidence for the attribution of individual beliefs if such attributions were supported by well-confirmed empirical laws. What could an intentional law be? What might some examples be? They give us a template:

> being in intentional state A is nomologically sufficient for being in intentional state B.

For such a law to provide leverage against the Quinian circle, we must suppose that it would describe a genuine discoverable regularity in the world, and not just express a disguised constraint on interpretation. For instance,

> being in the state of *believing that five is greater than three* is nomologically sufficient for being in the state of *believing that three is less than five*

is unimpressive as a candidate law precisely because, at least in as extreme a case as this, the rival suggestion is compelling that one just *wouldn't count* a person as being in the former state if one was not prepared to count him as being in the latter as well. The evidence for the left half of the "law" would be the same as the evidence for the right half. So let us swing, momentarily, to the other extreme, and imagine the discovery of intentional laws that bore *no* resemblance to such rationality-presupposing norm-approximations. Let us suppose, for instance, that it turns out to be a deep and unfathomable intentional law that

> being in the state of *wanting to go to Paris* is nomologically sufficient for being in the state of *deeming candy a better Valentine's Day present than flowers.*

The believer in intentional laws must say "Don't ask *why* people who want to go to Paris think this - as if it were somehow *rational* for this deeming to go with that desire. It just does - it's a law!" I have never seen a plausible example of such a candidate for an intentional law. It provokes suspicion in those of us who are skeptical about intentional laws that all the plausible examples we have encountered look awfully like stand-ins for their nearest normative neighbors. For instance, here is a generalization that is "confirmed by its instances, supports counterfactuals, and so forth":

> by and large *(ceteris paribus)* people believe there's a cow standing in front of them whenever they ought to believe there's a cow standing in front of them.

What then of Fodor and Lepore's example of the moon illusion, which they cite as a clear instance of an intentional law? As they say, there is a regular and reliable relationship between the moon's location and its apparent size, but why call it "lawful?" This regular relationship no doubt has some sort of ultimate explanation at the design-stance level. (Fodor and Lepore seem to forget that the intentional stance is one of three stances in my theory, which purports to explain the relations between them.) There is also no doubt a design stance explanation at a somewhat higher level of abstraction of the chess player who is suckered by knight forks - there is a sub-optimal bias in his heuristic search-tree - pruning dispositions, for instance. I never deny that we can take advantage of design stance discoveries; I just insist that they are interpreted *as* design stance implementations of intentional stance competences, and those competences are still describable only in rationality presupposing language.

The fixation on "laws" by philosophers is a mystery to me. I suspect it is leftover physics envy, but perhaps there is some other explanation. Does anyone suppose that there are laws of nutrition? Laws of locomotion? There are all sorts of highly imperturbable boundary conditions on nutrition and locomotion, owing to fundamental laws of physics, and there are plenty of regularities, rules of thumb, trade-offs, and the like that are encountered by any nutritional or locomotive mechanisms. But these are not laws. They are like the highly robust regularities of automotive engineering. Consider the regularity that *(ceteris paribus)* ignition is accomplished only by or after the use of a key. There is a reason for this, of course, and it has to do with the perceived value of automobiles, their susceptibility to theft, the cost-effective (but not foolproof) options provided by pre-existing locksmith technology, and so forth. When one understands the myriad cost-benefit trade-offs of the design decisions that go into creating automobiles, one appreciates this regularity. It is not any kind of law; it is a regularity that tends to settle out of a complex set of competing, *desiderata* (otherwise known as norms). These highly reliable generalizations are not laws of automotive engineering, nor are their biological counterparts laws of locomotion or nutrition, so it would be surprising if there were laws of cognition - intentional laws. And even if you decide for one reason or another to call such generalizations laws (or "c.p. laws") they won't play the anchoring role needed to break the Quinian circle, for attributions will *still* come down to questions of interpretation (analogous to: is this electronic device that remotely locks tile car door and disables the ignition really a *key?)*. Such norm-tracking design-stance generalizations abound in biology. The location of the mouth at the bow rather than the stern end of the locomoting organism *(ceteris paribus* - there are exceptions!) gets a similar style of explanation, presumably. And it will turn out that the moon illusion is a tolerable price to pay under normal conditions for other benefits to one's visual system.

This designer's-eye perspective on intentional regularities is familiar to anyone who has confronted the task of creating an information-processing system. One of the primary insights of computer science (or more specifically, Artificial Intelligence) is that you have to *arrange for* the mechanisms to generate (approximately) the regularities needed. Sometimes in the quest for such arrangements you may encounter deep, elegant principles or regularities that can be exploited, such as those that ground the enthusiasm for connectionism, but these cannot be Fodor and Lepore's "intentional laws" for they can be observed in entirely non-semantic, non-psychological contexts. Properly exploited, however, such regularities permit one to construct systems that approximate the sought-for intentional regularities in operation. When you succeed - close enough, anyway - you get to *mil* the states of the device you have made intentional states. This holism of attribution is obvious in AI, and I have never seen any reason to suppose that cognitive science harbors any other principles of attribution.

Ironically, then, my chief criticism of Intentional Realism is that it is unrealistic; it conjures up an imaginary cognitive science that has never existed, and never will exist - intentional surrealism. This can be discerned from a different perspective in Colin McGinn's essay, which shows, more clearly than ever before, how the truth about the normativity of intentional ascription gets warped by this commitment to Realism into the more luxuriant fantasies of logicism. So close and yet so far! McGinn's comments on the collapse of the problem of mathematical knowledge seem to me to be right on target:

> We can know logical truth only because we embody it, so if we didn't we wouldn't. We know the logical powers of the truth-functions, say, *only* by virtue of the (non-contingent) fact that the corresponding symbols in our heads exhibit causal features that map reliably onto those logical powers. (p. 93)

As McGinn says, "there is a strong sense in which we have to *be* logical if we are to know logic and mathematics." or anything else, I would add. But McGinn conflates the brain's *being logical* (having innately rational mechanisms) with the brain's *using logic,* and this move, which has a distinguished history, of course, creates a major distortion. McGinn notes (in agreement with "Intentional Systems,") that the ultimate explanation of our rationality must be evolutionary: "The reason our minds follow the causal tracks they do is that these tracks are designed to mirror logical relations; would-be minds that followed other tracks failed to pass evolutionary muster." But this is as true of fish minds and bird minds as it is of human minds, and fish and birds do not engage in the practice of abstract logical reasoning, any more than they engage in legal reasoning. The "chains of reasoning" McGinn speaks of are a minor, species-specific manifestation of the rationality evolution designed into us and other fauna. It would be impossible to build the artifacts of logic and language (the meme-structures - see the discussion of Dawkins and Dahlbom below) in a brain devoid of innate rationality, but we mustn't make the mistake of then trying to impose the structural details of those artifacts onto the underlying processes that make them possible. That this can seem to be necessary is itself an artifactual illusion of the meme of Realism: "What makes a belief the particular belief it is is the proposition believed, and propositions are precisely the proper subject-matter of logic." (p. 87).

As we have seen, two ways of reading this have been proposed: as a metaphysical fact about mental states (Realism) or as a constraint on ascription (Interpretivism). Take some particular mental state (the state of Jones when he sees the cow run by the window): one way of going asks: which belief *is* it? the other asks: which belief should we *call* it? The latter says, in effect,

> Don't call it the belief that *p* if you suspect [because of your knowledge of Jones' brain mechanisms, perhaps] that Jones will fail to appreciate some of the important logical consequences of *p,* for this would show that "belief that *p"* would *mismeasure* the actual

> state of Jones. Cast about for whatever proposition comes closest to capturing that actual competence of Jones in this regard. Suppose *q* comes closest; you cannot do better than to call Jones' state the belief that *q*.

The Realist view, alternatively, supposes that Jones's state *really has* some particular propositional content, so that there is a fact of the matter, independent of all these subsequent psychological states. Although McGinn cites Quine's discussion of radical translation, he apparently doesn't see it as what Fodor and Lepore would call a transcendental argument, for he sees no way of avoiding the Realist view of it:

> We must be charitable about other people's reasoning *because* their thoughts are inherently logically defined states. Without mental logicism as an ontological doctrine, logical charity looks like a kind of wishful thinking - or merely [!l a reflection of what the *interpreter* needs to assume about the world. The norms have to be written into the internal structure of the belief system if they are to condition its interpretation. (p. 89)

McGinn's discussion shows clearly that Realism makes sense only in cases in which we have chosen a language *and an interpretation of its terms,* for then we can arrange for particular, pre-identified sentences (whose meanings are already fixed) to be grouped together as axiom sets, etc. Pushed through the template of such a language of thought, *rational norms* get twisted into *logical laws:* "Nakedly stated, logicism says that psychological laws are logical laws" (p. 88).

McGinn imagines a field that postulates an innate logic module, equipped with transducers sensitive to structures isomorphic to the syntactical markers of logical form in formal languages. He calls this field computational psychology, but almost no cognitive or computational psychologist would endorse this vision (for a valuable early discussion, see Johnson-Laird, 1983). What McGinn has actually described is closer to the vision of John McCarthy and his school of logicists in AI. It is curious that for all the varieties of logicism that McGinn considers, he ignores the one that has most explicitly endorsed the identification he proposes between being rational and using logic. The base-level problem facing any intelligent agent is always: now what should I do?, and McCarthy and the AI logicists have supposed that all answers to this question should be generated by theorem-proving machinery operating on an axiom base of the agent's beliefs to date, with the "logical laws" built right in as features of the inference engine that feeds on the axioms and generates new lines of proof. There are many well-known practical or empirical difficulties with this approach (for a discussion of some of them, see my "Cognitive Wheels: an Introduction to the Frame Problem of AI," 1984), and there is also a conceptual difficulty brought to the surface by McGinn's suggestion that such innate logical machinery could account for human prowess in mathematics. If our rationality consisted in our being disposed to entertain sentences in the language of thought that followed logically from the sentences that we had previously entertained - if, in other words, our rationality consisted in our being *hard-wired consistent theorem provers* of one sort or another - we would be bound by the Godelian limitations that J. R. Lucas, and more recently Roger Penrose,[4] have taken to signal a *reductio ad absurdum* of the hypothesis of mechanism. Our capacity to "produce as true" or "recognize by intuition" the mathematical truths Godel's Theorem proves to exist could not be explained as the capacity to generate these truths as theorems.

McGinn speculates (in note 7) that perhaps I prefer what he calls an instrumentalist interpretation of mental logicism because I "can't see how a physical system could really, in itself and objectively, be subject to logical norms." On the contrary, I can see all too clearly how logicism runs and what the objections to this move are. And I have seen the alternative, which is really just a nudge of emphasis away from the claim McGinn makes. Why do people who believe *p* and *p* → *q* go on (as it were) to believe *q?* The fact that "such laws actually work" is to be explained, McGinn avers, by the fact that "the causal powers of the two premise beliefs accordingly mirror the logical powers of the premises themselves." The mild and undeniable way of reading this is as the claim that cognitive transition "tracks" sound inference: there will turn out to be an interpretation of the transitions between contentful states in the brain that "mirrors the logical powers" of the propositions used in the interpretation. It is when McGinn insists on a strict Realist interpretation of this explanation that he sets off down the wrong path, as we shall see more clearly when we consider the same issue in Millikan's essay.

6 Millikan on the Bio-mechanics of Representation: Why the Details Matter

The trouble with computational psychology as Fodor and Lepore, and McGinn, imagine it is that they have failed to imagine it in sufficient concrete detail. Principles that seem obvious so long as the investigation is sufficiently general have a way of evaporating when you get down to cases. What might computational psychology really be like? What would a realistic (as opposed to Realistic) theory of mental representation look like?

Ruth Millikan says: "What is needed is not to discover what mental representations *really are* but to lay down some terms that cut between interestingly different possible phenomena so we can discuss their relations." I would say *apparently possible* phenomena, for the deeper we go into the mechanical and ultimately biological details, the more hypotheses we have to discard. Millikan sketches some of the territory, and not surprisingly I approve heartily of most of it, since it elaborates a functionalistic view of content attribution of just the sort I have been recommending as an alternative to Fodor's atomistic Realism. But beyond that, it is full of novel and valuable observations. One of the most seductive metaphors in cognitive science, especially in its most theoretical and philosophical quarters, is the idea of a language of thought. So attractive is it that many have thought it has already begun hardening off into literal science. By taking the hypothesis of the language of thought seriously, as more than just the enabling slogan of various flights of philosophical fantasy, Millikan shows how many pitfalls still beset the notion. She ends up supporting and extending my skepticism about the hypothesis, while also finding and repairing some oversights and misdirection in my account.

> The "content" of an intentional icon is described by telling what sort of structure or feature would have to be in the organism's environment, for the icon to map onto by its mapping rule, in order for its consumer to use it successfully in the normal way .... (p. 100)

This supports and illustrates my fundamental point about constraint on interpretation. We are *bound* to ascribe content that makes it come out that there is a free-floating rationale in support of this design. When determining the semantic value of something, "we should refer ... to the general principles in accordance with which it is *designed* to be guided .... " In other words, there simply are no grounds for content ascription beyond those that account for *how well* the devices work at performing their functions when conditions are right.

Content can be defined with "considerable determinacy," she says, according to the following rule: "Consider the content to be that mapped feature to which the icon specifically adapts the *user(s)* of the icon. It is that feature which, if removed from the environment or incorrectly mapped, will guarantee failure for its *users"* (p. 101). She's right, I think. This is a good way of getting at the point. But isn't Millikan's view a species of "atomistic" Intentional Realism? After all, she attributes content to particular, individual "intentional icons" in the brain, one at a time - or so it first appears. But look more closely. Since the content (if any) of an intentional icon (if that is what the item you are looking at really is) depends as much on the "user" -part of the system as on its "producer," no item gets content independently of its role within a larger system. For instance, she notes that “A mental name exists only when its thinker has a competence to reiterate it in a variety of different correct and grounded *whole* representations ... " (p. 119). Typically, the larger system will only be able to play its "user" role if it includes within itself a variety (yea even a host) of other intentional icons, and then we're right back to "meaning holism." The individual structured items may be picked out "atomistically" but whether they are intentional icons, and if so, what content they carry, is still a holistic matter. In the limiting, simplest case there might be an icon whose "user" could use only one icon; precisely in these simple cases we have what one might call "degenerate" cases of intentionality - like the poor dumb thermostat that we call an intentional system for the same reason we call zero a number: denying it this status makes for worse boundaries than granting it.

Can't we identify an individual spark plug *as* a spark plug without looking at the rest of the engine? Sure. You can't explain the function of a spark plug without at the same time explaining its role within the larger system, but so what? We can detach these items and identify them in their own right. Now why couldn't we do the same thing with beliefs? Or, to use Millikan's term, intentional icons? A comparison with DNA is useful. It is mindboggling to learn that one can splice firefly genes into *plants* that thereupon glow in the dark; here is functional detachment with a vengeance. If there can be a glow-in-the-dark gene that can be moved from insect to weed, why not a belief-that-snow-is-white mechanism that could (in principle) be installed in the head of one who had never even heard of winter?

Suppose, then, you try to install the belief that snow is white in somebody who doesn't already have beliefs about the temperature of snow, its composition as frozen water, etc. If you don't bring all this information along (in some sort of package deal, let's suppose) then whatever else you may have done, the result will *not* be the installation of the belief that snow is white. It will be, perhaps, the disposition to assert "snow is white" (or its canonical translation in whatever language our patient speaks), and also, perhaps, to assent to questions like "Is snow white?" with affirmative answers of one stripe or another. This person still does not believe that snow is white, unless you have a new concept of belief you're selling. After all, *ex hypothesi* this person can't tell snow from bubble gum, hasn't any idea that it is a winter variety of precipitation, etc.

"'You're just changing the concept!' - are you not just making a move in the late and unlamented form of conservatism known as ordinary language philosophy?" Yes I am, and the move was not pointless, but only overused. We *might* want to change the concept of belief in this way, but then let's have the evidence that there really are units, elements, or whatever, that divide nature *this* way. Fodor and Lepore have none to offer, and Millikan's discussion of intentional icons suggests a rather different taxonomy of items.

But still, Millikan seems to suppose that the functional analyses on which she would ground her content-ascriptions bottom out at nice brute facts of the matter - facts of evolutionary history that answer the crucial questions about *the* function of one icon-system or another. Here is the one point of fairly substantial disagreement remaining between us. I have not convinced her that when she looks as closely at the details of her speculative evolutionary histories as she has looked at the details of her representational mechanisms, she will see that functional ascriptions are subject to the same norm-racked questions of indeterminacy of interpretation as intentional ascriptions. I recognize that I have yet to persuade many philosophers that they must take seriously my shocking line on how biology depends, in the end, on adopting the intentional stance towards the evolutionary process itself (that's one reason I am writing my next book on the idea of evolution), so for the time being I will leave matters conditional: if I am right, then Millikan's holism cannot be *quite* as "realistic about content" as it now appears.

Consider her discussion of the perennial question about what the frog's eye tells the frog's brain (p. 107). Is there a *definite content* (i.e. a proposition that is well expressed by a particular English sentence of whatever length) to what the frog's eye tells a frog's brain? Her novel suggestion is that while the particular intentional state - the "telling" - may not even imply (let alone be equivalent to) any definite English sentence or sentences, it might nevertheless be true that "English sentences can be constructed that will imply the truth of what any particular frog's eye is currently telling its brain." This strikes me as an ingenious way of responding to the legitimate demand to say something as definite as possible about tile content of this froggy telling without embroiling oneself in the problems of the mismatching "logical spaces" of English and Froggish. I noticed the problem in *Content and Consciouness,* section 10, "Language and Content," and disparaged the quest for *precision* but missed the idea that at least certain one-way implications might be established.

When Millikan attempts to make implications run the other way for people (if not frogs), I have my doubts. "It might still be," she says, "that *'core* information-storage elements in the brain'... would strictly entail the belief-ascribing sentences through which we customarily filter their contents" (p. 107). Strictly entail? How would this be understood? The particular information-ladenness of your visuomotor system at this moment - its content - may in fact contain elements such that, as Millikan puts it, various English sentences entail the truth of those elements, but I don't see how the inverse direction makes sense as a strict entailment. In "Beyond Belief" I imagined someone who had a "thing about redheads," a thought experiment designed to respond to this apparently well-posed question. It was poorly designed, apparently, since Millikan alludes to the example hut declines to draw the conclusion I had attempted to enliven: an organism might have a system in its head - we can be deliberately noncommittal about its structure and just call it a "thing" - that quite determinately and systematically contributed in a content-sensitive way to ongoing cognitive processes without that contribution being capturable as the contribution of a particular *proposition,* just because the system in question, and the supersystem of which it is a part, while content-laden, is not a sentential system.

Millikan's main point of disagreement with my view - or with what she takes to be my view - concerns the role I have given rationality. When she asks if, as my early slogan has it, rationality is the mother of intention, she has in mind a reading of rationality as logicality, starting down the same path as McGinn (and a host of others, to be sure). I think this obscures to her the continuum of process types that take us from brute, non-inferential tropism, through various inference-like steps, to full-fledged "rational thinking." One can start with the rationality of the design process that produces the mechanisms that do the work. If the thing is wired up rationally, it doesn't need to do any further "thinking;" it will do the rational thing. It doesn't have to (a) identify a state that has meaning; (b) figure out, using "rational thought," what it means; (c) figure out, using more "rational thought," what the implications are on this occasion of a message with that meaning; (d) compose a plan of action rationally justified on the basis of this determination; and finally (e) act on it. If it is rationally wired up, the mechanism will take care of all this. That is the sense in which I view rationality as the mother of intention, but when we look, with Millikan, at more specific ideas about how this wiring-up might be implemented, we find distinctions that can be drawn between processes that are in a stronger sense rational, by involving something more like formal inference.

Millikan introduces *mediate inference* as the touchstone of "representations" and gives a fine example; the critter that figures out that the lions are near its water hole by "pivoting" on the icon it has for its den. Her demonstration of the importance of a "middle term" is a good way of getting at what Fodor has long been insisting on: the importance of constituent structure in mental representations. Notice, though, that her way is accomplished without any presupposition of anything like a grammar (unless the mapping rules of map counts as grammar, which I doubt). She does it by semantic-level discussion alone. And the pivoting involved is not *inference* in a narrow, syntactic sense. Or at least I would urge that reading on her discussion:

> Suppose, for example, that, by combining information contained in perception of an object *seen* to be in a certain place, with information about that object as *felt* to be in that same place, one comes to believe, say, that a green apple is hard. Here the premises are percepts and the conclusion is a thought, but the motion turns on a middle term and it is inference. (p. 104)

Would she join me in extending the same account to such phenomena as the "pop out" of the red "disk" in the field of bagels, discussed by Churchland and Ramachandran? One of the "various jobs" of inference, she says, "is to combine with other icons to produce icons carrying new information." In this phenomenon I propose that the process in effect applies the "more of the same" inference rule, and since more of the same in this instance means more red, and since there is a circular outer boundary of red in the immediate vicinity, the system jumps to the conclusion that there is a red disk, and this conclusion then provides an odd-man-out premise for "pop out." Or is it a mistake to call these relatively early visual information processing steps inferences? This has been a hot debating topic within cognitive science for years - actually ever since Helmholtz.

Certainly "purely perceptual" processes can have the *logical effect* of an inference, including the inferences of identification that Millikan so illuminatingly concentrates on. One of my favorite examples is the experiment by Nielsen (1963), in which, thanks to tricks with mirrors, subjects *misidentify* a hand they see in front of them as their own - and the "conclusions" drawn from this misidentification include "feelings" of "pressure" preventing one's hand from moving where it is supposed to move *(Consciousness Explained*, p. 112n). Were it not for the mistake of identification, the phenomenology of "felt pressure" would not occur. Is this mistake an inference? If so, then inference is ubiquitous in perception; one need have no awareness at all of a passage of rational thought; the effect is experientially transparent or immediate. Presumably this can also be true in the case of the green, hard apple. Although one *might* become aware of - notice - that it was one's own hand (apparently) that picked up the apple, this is surely not required for one "to produce icons carrying new information." Millikan presumably has such cases in mind when she mentions inferencers who "rely heavily" on "tacitly presupposed information." She says that "a correct logical reconstruction of the inference" would require rendering this tacit content explicit. Indeed. This is the rule of rational reconstruction, but it is not a requirement on the mechanism that performs this quasi-inference.[5]

She says that the ability to make inferences is not at all the ability to recognize contradictions or inconsistency, and this can be supported in a way that may not have occurred to her. There are anomalies of perception in which people report, sometimes without wonder, contradictory perceptual states. Or even if they stumble over the contradiction and marvel at it, the effect persists.[6] For example, there are stimuli that generate "paradoxical motion," things that seem very clearly to move without changing their location, or wheel spokes that seem to accelerate their angular velocity of rotation without increasing their RPM. The tolerance of our brains for inconsistent conclusions is often striking.

Millikan's imaginative exercise illustrates a quandary I have faced myself. On the one hand, as she shows again and again, the details count. You are *not* entitled to assume that a "language of thought" system would work just because you think you can dimly see how a little bit of such a system might work. All sorts of dubious assumptions (e.g. about the typing of tokens, about distinctions that make sense for a human language but lapse in the context of brain mechanisms) lie hidden from exposure until you take on the task of describing the machinery of a language of thought with some serious attention to detail. The difficulties that her excursion encounters highlight the shaky position of those who think they can assume some such theory without worrying about details. On the other hand, however, these difficulties also show that the details count even more than she acknowledges! Her pioneering explorations are well stocked with acute observations, and yet in some regards she is going over ground that has been prospected for years by researchers in other disciplines, going back to such classic AI papers as William Woods' "What's in a Link?" (1975). And even these AI investigations, more constrained by mechanical detail than most philosophical explorations of the issues, exhibit the vulnerabilities that Church land and Ramachandran diagnose in my own top-down ruminations.

The trouble with thought experiments such as Millikan's (and I've been fond of them myself) is that they are unreliable. It is all too easy to imagine them partially, and on the basis of that imagining, conclude either "Oh, then it's possible to do it this way - just in a more complicated version" or "So, you see, it isn't possible to do it this way." But think of how similar this epistemic situation is to one's predicament when having just seen a magic trick. You think about possible mechanisms for awhile, and finally (usually) you give up - and if you didn't know better you'd declare "It just isn't possible. There is *no way* that card could have gotten into that box (or whatever)." And then when the trick is revealed to you, you discover a simple little loophole that you have failed to close, a simple hard-to-think-of alternative. God does not need to be malicious to have exploited lots of these loopholes in nature. So while simple mechanical examples, such as Millikan's sliding board generalizer, may serve a valuable purpose in opening our eyes to (apparent) possibilities, they should be treated with caution.

7 Do Bats Have a *Weltanschauung,* or just a *Lebenswelt?*

Kathleen Akins' essay illustrates how easily I, and others, have had our vision narrowed by ignoring these other perspectives. I have held that adopting the intentional stance towards such creatures as bats is not only harmless but positively useful anthropomorphism. It permits one to "set the specs" - to describe the information-processing tasks - in advance of hypotheses about specific mechanisms. But what if that very setting of the specs already seriously misdescribes the requisite competence, by imputing - along with the bat's beliefs and desires - a human-centered ontology, or other features of our human conceptual scheme that are gratuitous sophistications? This is a more radical version of the hypothesis that the "logical space" of Froggish is too different from our logical space to permit anything like translation. This is the hypothesis that bats (or frogs) might not really have enough of a logical space of concepts to have anything worth calling a "world view" at all.

For instance, Akins asks what evidence would justify the attribution of the concept of an object to the bat. She points (as does Millikan) to the important operations of re-identification and re-location. Do bats *keep track of* individuals? I don't lose track of the individual water molecules I drink; I don't track them at all. I *can* track a particular grain of sand in my shoe, but only under rather special circumstances. The fact that the issue of re-identification for *it* can come up does not mean that I lose track of the other grains on the beach: tracking them is not (except under weird conditions) an option for me at all. "Sand" is a mass noun for me, like "water." If, Akins suggests, bats don't have any use for the concept of an object (a particular object), then they could hardly have beliefs *about* objects. Moreover, without exemplars on both sides, they would have no use for the mass-noun/count-noun distinction. This surprising application of Strawson's metaphysical analysis of the concept of an individual to problems in neuroethology shows that there really is work for "neurophilosophy" to do.

Setting aside discussion of all the discussable details for another occasion, I must comment on two main themes in her essay. First, it demonstrates better than ever how misleading Nagel's wonders are about the *intrinsic* properties of bat experience. At least a great deal of what it is like to be something actually concerns such informational matters as the way opaque objects occlude things, the way distant objects are fuzzy and indistinct, and so forth: "representational problems - *not* problems about inaccessible subjective facts or the intrinsic properties of the experience or the phenomenological 'feel' of a subject's experience." By a careful study of these informational matters, we can close in on the question of what there *could be* in the bat's world. Can we close in all the way? Won't there be an inexplicable residue? Akins' discussion at least suggests a way of shifting the burden of proof: show us, Tom Nagel, *any* interesting subjective fact that will not eventually be pinned down by this process. (I play Burden Tennis when the conditions call for it; I just don't think it's the only game in town.)

That's the positive theme. The negative theme is more striking: there might not be anything at all that it is like to be a bat. "Nagel inadvisedly assumes" that a bat must have a point of view, but, Akins claims, if the bat is well designed in a cost-effective way, this may be gratuitous. Recall the amazing success of the Oak Ridge uranium-enrichment plant during the Manhattan Project. Thousands of workers spent their days controlling the highly complex process, making all manner of timely, appropriate adjustments in the face of a huge repertoire of conditions *and they hadn’t a clue what they were doing!* Not only did they not know they were helping to make an atomic bomb; they also didn't know they were enriching uranium, or enriching anything. Never was the motto truer: "Don't ask me, I just work here!" The bat's imagined wingflapper is similarly positioned; it doesn't have to know it is controlling the flight of an insectivorous mammal, or controlling flight at all. "The same kind of story could have been told from the vantage point of some other cortical area . .. [i]n each case, a different story would have unfolded - another 'subjective world' - a story in accordance with the external information available at that particular site" (p. 153). And where in the brain, we feel tempted to ask, does that bat *itself* reside? What makes us think there is any good answer to this question? There are many parallel stories that could be told about what goes on in you and me, and what gives (roughly) one of those stories pride of place at any one time is just that it is the story you or 1 will tell if asked! If the creature in question isn't a teller - has no language - then the supposition that one of these stories is privileged, in that it would tell what it is actually like to be a bat, dangles with no evident foundation beyond tradition.

It is for just this reason that 1 insisted, in *Consciousness Explained,* on beginning with human consciousness first, for there we do have a variety of robust and traditional anchor-points for our intuition that a single unified perspective on the world attaches - one-to-a-customer - to each human being. Once we have clarified the sources of *that* intuition, and seen its very real limitations, we can go on to explore the case of consciousness in other animals, less sure of the optimistic assumptions Nagel takes for granted. From the fact that the bat's sonar information is fed directly to the wingflapper without having to pass through some imagined bat-headquarters, one cannot directly infer that it doesn't *also* contribute to a higher level, of course. After all, we human beings can both walk and become conscious of our walking. The new question about bats then becomes: do bats, lacking language and all that comes in the train of language, *have* a higher level, and if so, what is its role in bat life? When we have answered that question - and I see no reason why that question can not be answered definitively from the third-person point of view - we will know if it is like anything to be a bat, and if so, what it is like.

8 From Biology to Society

The gulf between bats and us is not the gulf Nagel describes, but it is real enough, and Akins expresses from the bats' side much the same criticism that John Haugeland and Bo Dahlbom express from our side: my attempt to span the gap with the intentional stance is at best misleading, at worst a major theoretical error. Resistance to my unification proposal has been around from the beginning, but usually in the unprepossessing form of declared intuitions to the effect that lower animals (and robots) don't *really* have beliefs (not *real* beliefs, not really), a scruple that was easy to shrug off since my position was that whatever we decided to call them, there were informational states the having of which permitted both bats and bartenders to be predicted, manipulated, explained from the intentional stance, and this was what was theoretically important. But the versions of skepticism expressed by Akins, Haugeland, and Dahlbom cut deeper: we are social; bats (and chessplaying computers) are not, or at least not in certain ways that are critical for distinguishing between different *theoretically important* sorts of informational states. While Millikan and Akins concentrate on the confusions that can be engendered when we blithely impose what are essentially social (e.g. linguistic or even metaphysical) categories on biological phenomena, Haugeland and Dahlbom concentrate on the illusions that can arise when we ignore the social dimension and try to biologize all phenomena of human cognition and consciousness.

As Dahlbom notes, it is not that my position has left no room for drawing these important distinctions, but rather that I have failed to exploit the room I had created - in the distinction I had drawn but largely neglected between beliefs and opinions (in my technical sense of the term). My chief regret about the contents of *The Intentional Stance* is that I did not include in it a chapter on opinions, renewing and extending the claims I advanced in "How to Change your Mind" in *Brainstorms.* Now is an opportunity to repair some of that oversight, with the aid of some concepts drawn from Richard Dawkins: the extended phenotype, and the meme.

Both Haugeland and Dahlbom note the necessity of norms, but, unlike McGinn, Millikan and me, they insist that these norms are social at their foundations. But how could this be? How could social norms - and the holding to them that Haugeland describes - become established except on the shoulders of biological norms, of the sort that Millikan has explored? In "Intentional Systems" (1971), I argued that the concept of belief "had a normative cast to it" that was difficult to capture, and went on to propose that far from its being the case that the norms of belief were in any sense *voluntarily followed* (under the sort of mutually acknowledged carrot and stick that Haugeland describes), the norms of belief were constitutive at a lower level (the level so well described by Millikan). Only in creatures that *already* were error-shunners or truth-trackers (whether they knew it or not) could the foundations be laid for more self-conscious and sophisticated cognitive transactions.

And "transactions" is, I think, *le mot juste* here. Above the biological level of brute belief and simple intentional icons, human beings have constructed a level that is composed of *objects* that are socially constructed, replicated, distributed, traded, endorsed ("I'll buy that!"), rejected, ignored, obsessed about, refined, revised, attacked, advertised, discarded. Dawkins has given us a generic name for these things - memes - and what I have called opinions are a species of memes: sentences on whose truth a person has, in effect, made a wager. Non-human animals have nothing to do with memes, and hence their minds are vastly simpler than ours, as Akins and Haugeland say, looking across the canyon from opposite sides. One might say that, whereas animals (including human animals) have brains, memes turn brains into minds, arming them with intellectual tools, as Dahlbom says, and *thereby* extending the human phenotype into the larger world of civilization. A human being denied access to memes (by deafness or isolation, say) is like a hermit crab exiled to a world without vacant shells to appropriate, fundamentally incomplete and ill provisioned to face nature in the ways it was designed to do. As Dahlbom says, we are to a large extent artificial ourselves, even if our bodies aren't. We're Catholics, atheists, electrical engineers, licensed drivers, computer-literate, polyglot. Millikan provides a valuable step here: human beings don't just have the benefit of evolution-designed intentional icons; they have the benefit of evolution-designed capacities to *acquire* further intentional icons, and a host of other useful objects, from the social world they grow up in. They can also *abandon* these objects - as Anthony Kenny abandoned his priesthood - in quite sudden revolutionary changes of "psychology" that are unknown in the animal world, where "one-shot learning" of even the simplest "facts" is exceptional.

So I can agree with Dahlbom that it is a mistake to say that thinking occurs (just) in the brain; it is the same mistake as asserting that sexual display occurs in the body (ignoring the bowerbirds' bowers, and the teenagers' hotrods) or that feeding occurs in the mouth (ignoring the spiders' webs and the farmers' plows). But I disagree with his suggestion that explanations of the social world and its objects are discontinuous with or even incompatible with explanations of the biological world. There is indeed a tension between the two perspectives (manifest in Dawkins' question-begging defense of the memes of science), but these two perspectives can be unified. The key to unification lies, I think, in *denying* what seems obvious to Dahlbom:

> If one wants to argue, as Simon clearly does, that the idea and methods of an artificial science are importantly different from those of the natural sciences, even if we are only beginning to appreciate how, then one should be careful not to describe evolution by natural selection as a process of design. (p. 175)

On the contrary, what has to give is the conviction that biology is a "natural science" in the way physics is - with "laws of nature" and "natural kinds." I agree with Dahlbom that there are important differences between the artifacts of engineers and authors on the one hand, and artifacts created by the process of natural selection on the other (see my "Artificial Life: A Feast for the Imagination", (1990), and "Cognitive Science as Reverse Engineering: Several Meanings of 'Top-down' and 'Bottom-up,'" forthcoming), but he has not convinced me that these outweigh the similarities.

A pivotal event in provoking this discussion was my choice of the chess-playing computer as the leading example of an intentional system. Both Dahlbom and Haugeland draw attention to the undeniably important fact that, at least to date, chess-playing computers have existed in a severely truncated version of the real chess-playing environment: they are entirely unequipped to confront such options as cheating, tipping over the table, deliberately losing in order to boost the morale of an old friend, giving up the game in favor of waterskiing, etc. And lacking such complexities, their intentionality is seriously attenuated. But is this really so important? I chose chess-playing computers because they were not science fictional, a qualification always worthy of philosophers' consideration, but not always of overriding significance. What if I had settled instead on firewood-gathering robots? Gone are the subtleties of institution-borne norms (until the robots start venturing onto *your* land for wood for *my* fire), and the concreteness of the objects "considered" and manipulated by the robot would be exemplary. Otherwise, the application of the intentional stance to them would go through in the same fashion, so far as I can see.

But perhaps I cannot see as far as my critics. Both Haugeland and Akins can be read to be claiming something that strikes at the heart of my idea about intentional systems: some intentional systems don't exhibit *intentionality* at all: not all God's critters got ontology, you might say. For Haugeland, the ingredient missing in non-human (and robot) intentional systems is social; in the absence of the memosphere - in the absence of social norms and all that attends them - there is no ontology. This is not the trivially true claim that ontology, as an academic sub-discipline of metaphysics, would not exist unless there were universities, books, etc., but the startling claim that animals (or computers) without social norms "have" no ontology; nothing exists "for" them. The idea that one can speak of, imagine, argue about "the world of the bat" or "the objects that dolphins can conceive of" is a big mistake! I read Akins' essay as consistent with Haugeland's stronger claim, but she may have other complications in mind. If bats don't have an ontology, they lack *something* that is requisite, but she doesn't commit herself to Haugeland's claim that the social realm is the right place to look for it. Now *if* one fails to make the belief-opinion distinction, and hence tends to think of animals' beliefs as sentences-deemed-true by those animals (sentences in their languages of thought), then one will have a vision of what it would mean for a bat or a dolphin to "have an ontology" that does indeed deserve suspicion. (As I once put it, do polar bears *believe that snow is white?* Not the way people do. Polar bears couldn't *think* that snow is white).[7] As Dahlbom says, *thinking* is an artificial, not a biological, process.

Or, to put the point from Dawkins' perspective, the memes that compose us interact in ways that provide a new layer of complexity on the biological substrate. Here new puzzles confront us. When should we say that a particular "virus of the mind" *exploits* the other memes that it encounters, and when should we say that *we*, masters of our ideas, *evaluate* and *choose* the doctrines we will live by? The idea that there is a *foundational* distinction to be drawn between these two poles is another element of our Cartesian legacy that we must learn to do without. In the context of computer viruses, Dawkins distinguishes the category of Trojan horses, which are not *directly* self-replicating code-strings, but rely on people's choices to get themselves replicated. These viruses are mere like pure memes, since the "alternate host" of a human mind is required, but the difference is a matter of degree, in the end, for we choosers have an incomplete knowledge of what lies inside the memes we favor.

Every meme is potentially a Trojan horse. What is "visible" to us, the (willy-nilly) selectors of memes, is one thing; what we thereby actually end up replicating is another - an abstract functional property of our nervous systems that has largely undreamt of powers of combination, subversion, and exploitation. We don't consciously take on all the logical implications of the ideas we adopt, and we can't assay in advance the practical effects of adopting them. Unlike natural selection, we are *somewhat* foresightful choosers of memes, but we should not underestimate the myopia and astigmatism created by the other memes that are our accomplices.

Dawkins overstates the case for the power of scientific rationalism. In his own field of evolutionary biology, not all good scientific ideas spread and not all bad ones are extinguished (Richards, 1987; Cronin, 1991). His own memes have had a rocky career in some niches, and some of those he has combated have had momentous (if, one hopes, short-lived) population explosions. Fortunately, however, one of the dominant meme-families that compose us - not all people, but those who have been infected by Western rationalism - has the property of being an omni-critical filter of other memes (and even, as Rorty's essay illustrates, of itself).[8] This is the artificial process of deliberately rational thought, a phenomenon idealized by the deductive procedures of academic logic (a family of meta-memes which includes diagrams and methods that *could not* be accomplished without language and even writing systems). It is not directly part of our genetic inheritance, as McGinn imagines, nor do its regularities depend directly on "laws of nature" that mirror the "laws of logic." *That* is the implication of Simon's claim that it belongs to the sciences of the artificial. The potentially unlimited spiral of meta-evaluation that it makes possible (and attractive) is what distinguishes human psychology from that of all other animals.

I am not yet persuaded, however, that attenuated concepts of ontology (and the other sophistications of the human world) don't apply fruitfully to the simpler inhabitants of the planet, but I do grant that the issue has been opened up by Akins, Haugeland and Dahlbom in a way that commands attention, and I have scarcely begun to answer the challenges they raise.

9 Rorty and Metaphilosophy

Dick Rorty is right in his surmise that I like doing metaphilosophy less than he does, perhaps because I'm not nearly as good at it as he is. His historical scenes are delicious: Hobbes and Gassendi staring incredulously at the Cartesians trying to carve out a secure sanctuary for consciousness, Darwin to James to Dewey (the anti-essentialism double-play combination), and the comic suggestion that my modest proposals to Nagel are like Hegel suggesting to Kant that he might, after mastering Hegel's new theory, become able to describe the thing-in-itself. I particularly commend his analysis of the roles that Wittgenstein, Ryle, Sellars, and Davidson play in this drama. If only all the other participants in contemporary debates saw it his way - our way.

He is also a remarkably fairminded and farseeing referee in the sport of Burden Tennis, Rorty's account of the issues that divide Nagel and Searle from me is the best I have encountered. Though these two chief opponents of my views have not contributed directly to this volume, their views get a fine airing by Rorty, I think. The fact that he comes down on my side does not make it less true that he has seen what they have been up to with more sympathy and insight than I have. His strategic comments about which side "wins" under which conditions are always apt (except, as noted above, regarding my claim of neutrality), but slightly in tension with his can for an end to what Janice Moulton cans the "adversarial culture" of philosophy. How can somebody who is so keen on the Sport of Philosophy - who can chide me for "misplaced courtesy to a half-defeated enemy" - issue his can for us to lay down our racquets and have a nice conversation? Robert Nozick once joked about philosophers so combative that their ideal argument would be one so powerful that it set up reverberations in your brain: if you didn't accept its conclusion, you'd die! (1981, p. 4) Rorty suggests that neither Nagel nor I can come up with such a killing argument. "Both their spades are turned," he says, because "it is not clear ... that there is any compelling reason" to change one's mind about such ultimate matters. I'll settle for that. The reasons I've given are only supposed to be compelling to those who prefer understanding to mystery, and there is actually quite a lot to be said in favor of standing fast for mystery. It doesn't move *me,* but so what? I certainly don't think Nagel is committing some (fatal!) logical gaffe in maintaining his position; there is even something wonderful about it - Horatio at the bridge and all that. If he would rather go on believing in intrinsic and ineffable properties, then he will have to forgo the fun of being in on the kin when we knock off the mind-body problem, but some people are squeamish about myth-murder. (What I can't win by honest argument Rorty encourages me to capture with a rhetoric that makes "verificationism seem glamorous, exciting, and exactly what the age demands." I'm trying, I'm trying.)

Rorty eggs me on in this game, largely approving of the plays I have made to date, but urging me to be more ambitious, more radical, more dashing as a metaphilosopher, than I have been willing to be. Is it ungracious for me to resist such an invitation? The program that would permit me to diagnose Nagel's "ambition for transcendence" as a "tender-minded yearning for an impossible stability" is almost irresistibly attractive, and contains much that I heartily approve of, but I see both general and specific problems with the radical positions Rorty thinks I should hold.

First a specific problem. Once we move beyond the scheme-content distinction, we will have no use, he says, for the intentional-real distinction. "We shall just can them 'objects' *tout court.*" I don't think so, or at least we will still need a more modest distinction between perhaps *mere* intentional objects and other (intentional?) objects. Consider the gold in Fort Knox. One of the interesting facts about it is that its role in the world economy does not depend on its actually being there, but just on people's *believing* that it is there. For instance, much of Belgium's vast gold reserves have been stored at Fort Knox since the beginning of World War II, and it will stay there indefinitely, in all likelihood. If it could somehow be secretly moved to another location, without any leaks to the Belgians or other interested parties, the carefully created intentional object, *the Belgian gold reserves in Fort Knox,* would still be intact to do its stabilizing work, while the real object might be scattered. Yes, of course, there will be circumlocutions that can do justice to such distinctions without reintroducing the old metaphysical divide, but won't they, in the end, let us go right on saying almost everything anybody wanted to say in terms of the intentional-real distinction?

Finally, my general problem. I am not convinced that what is true about consciousness is true about everything. Rorty suggests that my attack on the Cartesian Theater is just one skirmish in the broader attack Oil the "captivating but trouble-making picture of human inquiry: 'Penetrating the Veil of Appearance'." I do not agree. The well-regarded distinction between Orwellian and Stalinesque tamperings with the evidence lapses, I claim, in the ultra-close quarters of the brain, but in other quarters my allegiance to it is unalloyed. There *is* a fact of the matter about whether Oswald and Ruby were in cahoots, even if it forever eludes investigation. It is not *just* a matter of which story plays well to which audiences, and I would say the same about all the usual quarries of empirical investigation. It is only when we confront the Observer, that, as Kant put it, the distinction between the "for me" and the "in itself" breaks down. As I said at the outset, I'm actually still quite conservative in my adoption of Standard Scientific Epistemology and Metaphysics.

This essay has already overstayed its welcome; doing justice to the many points raised in the essays in this volume would require an essay at least twice as long. It would not be safe to assume that I tacitly agree with all the objections and interpretations I have not discussed. Even more important, no conclusions should be drawn about my sense of the relative importance of these essays or the difficulties they raise for me by counting the pages of discussion I have devoted to them. I discovered problems too minor and too major to deal with in this setting. There will be other occasions, and I hope to be better prepared for them, thanks to the help of these and other constructive critics.

NOTES

[1] "Every type of inquiry sets questions which probably only it can solve: how come the sciences can provide solutions, and indeed determine so well what counts as a solution? How do the sciences manage the trick of being, in a way, self-authenticating, internally determining what shall count as true or false?" Ian Hacking, *London Review* May 28, 1992, p. 6 (review of Bryan Appleyard *Understanding the Present: Science and the Soul of Modern Man*. London: Picador, 1992).

[2] In my review of Roger Penrose's *The Emperor's New Mind*, in *The Times Literary Supplement*, I 989, I called this orthodoxy to which I subscribe the Cathedral of Science. We are currently experiencing a wave of science-bashing. In England, Bryan Appleyard's recent book (see note 1 above) has the following message on the dust-jacket: "This is an emergency, Appleyard writes, because we must now find our true nature before science crosses the final frontier of the human self." In the light of this contemporary attitude, my relatively uncomplicated "scientism" may appear particularly shallow and unphilosophical. So be it. I have not been able to find a more secure bedrock from which to conduct my investigations.

[3] Fodor and LePore say "Well, either ecology does underwrite a notion of function or it doesn't," but elsewhere Fodor has recently opined that the distinction between *selection* and *selection for* cannot be reconstructed (Loewer and Rey, 1991, p. 296). This suggests that he has joined the puritans, and decided that evolutionary biology does not underwrite a notion of function after all.

[4] For an analysis of Penrose's error, and suggestions about the nature of the algorithms our brains actually run, see Dennett, "Murmurs in the Cathedral" (1989) and "Betting your

Life on an Algorithm" (1990). In Haugeland's essay, the important discussion of whether recognition could be a matter of rule-following invites the same response (with which I think he would agree): our capacities to recognize things could be based un algorithms without being based on algorithms *for recognizing.*

5 Cf. McGinn, who claims that "since pains and the like are not even links in chains of reasoning" he will set them aside in his discussion. This is a clear case of a philosophical artifact in the making. Step one: by "draw and quarter," force appeals to rationality into appeals to "logical norms" or, better, "logical structures" (and around the bend comes the language of thought). Step two: shelve off pains and the like, since they don't figure in "chains of reasoning" (grounds that would not make sense without the push in step one). Step three: notice that only a "substantial core" of a given subject matter has a logical nature, which then opens the door for postulating "modules" that handle the logical part.

6 There is even a neuropathology of misidentification. In *jamais vu,* one's perceptual experience is pervaded by a sense of unfamiliarity, or lack of affectual tone that can lead to amazing denials of identification. In the Capgras delusion, one becomes convinced that one's spouse or other loved one has been replaced by an identically appearing impostor. (For a sober account and a plausible theory of the mechanism, see Young, 1990.) Shades of Twin-Earth!

7 Cf. Norman Malcolm's (1972) distinction between *thinking that p* and *having the thought that p,* discussed by me in "How to Change your Mind" (in *Brainstorms).*

8 The toxicologist Simon Wolff (1992, p. 44) writes: "Science is no more and no less than one way of systematising knowledge. And modem science is a successful and lasting way of compiling knowledge because it is evolutionary, flexible and internally consistent." But how can it be evolutionary and not be a virus? Whence cometh this internal consistency? It is just that the meme for internal consistency is a part of it.

REFERENCES

Cronin, H. (1991) *The Ant and the Peacock.* Cambridge: Cambridge University Press.

Johnson-Laird, P. (1983) *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness.* Cambridge: Cambridge University Press.

Loewer, B. and Rey, G. (eds) (1991) *Meaning in Mind: Fodor and his Critics.* Oxford: Blackwell.

Nielsen, T. I. (1963) Volition: a new experimental approach. *Scandinavian Journal of Psychology*, 4, 225-30.

Nozick, R. (1981) *Philosophical Explanations.* Cambridge, Mass.: Harvard University Press.

Richards, R. J. (1987) *Darwin and the Emergence of Evolutionary Theories of Mind and Behavior.*Chicago: University of Chicago Press.

Wolff, Simon (1992) Science as the dustbin of hope. *New Scientist,* May 30, p. 44.

Woods, W. (1975). What's in a link? In D. Bobrow and A. Collins (eds.), *Representation and Understanding,* New York: Academic Press.

Young, A. (1990) Accounting for delusional misidentifications. *British Journal of Psychiatry,* 157, 239-48.