15

With a Little Help from My Friends

Daniel C. Dennett

With a Little Help from My Friends

Tom Sawyer's trick of getting his friends to vie for the privilege of helping him whitewash the fence made a big impression on me when I was a boy, and I have had great luck emulating Tom over the years. Nobody could afford to hire the talent that has labored on improving my theories and fixing my arguments, and never has my sense of this been stronger than at the St. John's, Newfoundland conference. Philosophy is often done in an atmosphere of one-upmanship and triumphant refutation, and nothing about our discipline damages its reputation in neighboring fields more. But here is a collection of essays that are often deeply skeptical of my positions, and boldly critical of my arguments, but always presented with an eye on how to get the weaknesses replaced by strengths, the problems repaired. They go beyond what I have said, beyond what I have tried to say, beyond what I understand even now, after reading and rereading the essays and writing and rewriting this commentary.

I have come to grips with some of their proposals and arguments quite well, I think. I have seen what was up and what to make of it. Sometimes I agree; sometimes I don't, and say why. But there is a convergence of attention by several authors on ontological questions—on realism of one sort or another—that I have still not been able to sort out, not because I think they are all wrong, but because I am tantalized by them all. This is the major shortcoming (I'm pretty

sure) of this commentary. I ought to have been able to digest and put into a single perspective the various constructive suggestions regarding ontology coming from (in alphabetical order) Kenyon, Lloyd, Ross, Seager, Thompson, and Viger, but I can't get such a consilience to settle down yet. I can only list the points of convergence and invite the readers of this book to propose further unifications.

The order of my commentaries is based on an obvious principle of building: Wherever the comments on one essay seemed to me to benefit from following the comments of another, I put them in that order, until all were in a single string. Not surprisingly, the one essay by a nonphilosopher is the one that seemed to need the least philosophical wind-up before I could comfortably deliver.

Evolution: Crowe and Dumouchel

Timothy Crowe shows that the problems of communication among evolutionary theorists continue to be severe, and that my own interventions have not achieved the clarifications I had sought. This is frustrating, but let's just try to fix it, locating the points of misunderstanding and ignoring the issue of who's to blame. Crowe characterizes my overall aim in *Darwin's Dangerous Idea* (DDI henceforth) as the attempt to show that "Natural selection . . . plays an essential role in the understanding of every biological event at every hierarchical level, from the creation of self-replicating macromolecules to evolutionary lineages." It all depends on what you are trying to understand about a biological event, of course, but yes, I do think that natural selection plays an essential role—some role or other—at every scale, though in different ways at different scales. But once we get down to cases and ways, I rather doubt that Crowe would disagree with me, since mine is not, I think, an extreme view.

Crowe breaks down his interpretation of my position into three aims, and argues that I don't achieve two of them. Batting .333 is fine in baseball, but I aspire for better. He accepts (1) my characterization of skyhook-free evolutionary explanation, but thinks I fail (2) "to show that natural selection equals biological engineering driven by a foolproof, gradual, step-by-step, substrate-neutral, algorithmic design process." He uses his own research with African

guineafowl as a telling example for showing where, by his lights, I've gone wrong on this second quest. He also finds fault with my attempt (3) "to prove that natural selection is a 'universal acid' that has effects and applications well beyond biology," including the development of human culture, but since his remarks on this score are so brief, I will have little to say about them.

I'm interested that the one success he grants me, defending a skyhook-free vision of evolution, he sees as uncontroversial, perhaps because he diminishes it from the outset: "I know of no evolutionary biologist, including Gould, who disputes the central importance of natural selection and its various cranes in the process of adaptation." My claim, however, is not the bland truism that cranes are "of central importance" but that *there are no skyhooks,* and to this day Gould has declined a number of invitations to acknowledge this point, perhaps because he is still uncertain just what a skyhook is. This surprised me, by the way. I had thought when I introduced the skyhook/crane distinction to Gould, some years before DDI came out, that he would grant me that point, and then go on to articulate his objections to (his version of) neo-Darwinism on that shared foundation. But he resisted. So Crowe's welcome attempt at ecumenical agreement on this first point doesn't quite get the job done. But leaving that issue aside, let's turn to the main topic of Crowe's essay.

First let me break down (2) into its constituents. Natural selection is

(a) Biological engineering;

(b) Algorithmic;

(c) Gradual, step-by-step;

(d) A substrate-neutral design process.

These are, I think, distinct claims that do not have to stand and fall together, though I will defend them all against his criticisms.

Biological Engineering

Here there has been a miscommunication of my aim, since I agree with the points he makes about the process of natural selection. In

particular, I agree, indeed insist, that natural selection does *not* involve forward-thinking, goal-directed R&D, but I claim that that does not mean that it does not involve R&D. I can see why one may naturally assume that R&D is *by definition* forward-looking and goal-directed, but it is precisely my aim to resist this definition and insist that the process of natural selection is importantly *like* human R&D *in spite of* not being forward looking and goal-directed. Perhaps, then, we are just differing on the aptness of the R&D or engineering label, as he surmises. But I would still disagree with a point of contrast he urges on us: "The cranes of natural selection lift the fitness of individual organisms, demes, and populations, *rather than* [my emphasis] developing biological design-features that are in reality effects of that upliftment." I don't understand the "rather than"; I would replace those words by the word "by." What is the issue, then? I am not at all sure. Perhaps it is this. Do we view the long-term survival of the eagle (individual lineage, demes, populations) as a by-product of the natural selection of its excellent eyesight and wings (among other adaptations), or do we view the emergence of its excellent eyesight and wings (and other adaptations), as a by-product of the long-term survival of the eagle (demes, populations, etc.)? Shouldn't we view both as the emergent effects of a long history of natural selection? I think this is a nonissue, and I have certainly not claimed that there is any foresighted design project of creating better wings *so that* eagle lineages might prosper.

Algorithmicity

So much for (a); let's consider (b), algorithmicity. Crowe takes his research on guineafowl evolution to demonstrate that I am wrong about the algorithmicity of natural selection. He argues that the diversity of phenotypes among the subspecies of mutually cross-fertile guinea-hen varieties is due to

> an incidental, accidental, indirect, even maladaptive consequence of normal adaptation, a process that falls squarely within the realm of historical contingency. Furthermore, contrary to Dennett, such situations show that natural selection does not time and again produce the guaranteed results generated by an algorithmic process.

But I see nothing in his account of the guineafowl that challenges my view of evolution. Evolution by natural selection is often a noise-amplifier, an exploiter of "frozen accidents," a tolerator of drift, not only in the case of Kimura-style accumulation of random mutations that are not expressed phenotypically, but also of phenotypic don't-cares, features that may vary even into maladaptive regions at least for a while. Crowe's misunderstanding here is particularly frustrating, since I anticipated just this misreading in DDI, and went to some lengths to forestall it with a variety of examples (52–60). I noted that the elimination tournament algorithm, as a sorting algorithm, is much more akin to the algorithms of natural selection than, say, long division is, and then I gave several examples of tournament algorithms that involved mixtures of chance—massive contingency, luck—and skill. My coin-tossing tournament is unadulterated 100% contingency, and yet it *is* an algorithm, after all, guaranteed to produce a winner, every time. Or consider the tennis tournament supplemented by Russian roulette: The winner in each round puts a revolver to his head, spins the chamber and pulls the trigger. If he's lucky, he advances to the next round; if not, his just-defeated opponent advances instead. Again, this algorithm is guaranteed to produce a winner, but it is also a contingency amplifier; it doesn't guarantee at all who (or what) will win. This contrast between "selective determinism" and "historical contingency" (to use Crowe's terms) is a remarkably persistent red herring—and I hasten to add that he is far from being alone in resisting the many corrections Dawkins and I (among others) have issued on this score. For some reason, it doesn't matter how many times you say "there is no conflict between algorithmicity and contingency" or how many ways you illustrate the point; some people are determined (!) to interpret algorithmicity as "selective determinism." Maybe there's a gene for it! (Joke) Seriously, I would like to know what Crowe makes of the fact that I cheerfully embrace his research as grist for my mill. This must come as a surprise to him. What in my writing led him to think otherwise?

So far as I can see, Crowe's research on guineafowl exemplifies circumspect adaptationist theory at its best—and he himself welcomes that prospect in his penultimate paragraph. The role of

"dynamic creation and reconnection of habitat islands" in creating geographical isolates, together with drift within the isolates, may suffice to explain the differences in the absence of any more direct or powerful adaptational forces, favoring the different arrangements. But then he offers a further argument, which I am dubious about:

> Because of their central importance to successful interbreeding, the components of SMRSs [specific mate recognition systems] should be under extremely strong stabilizing selection to "resist" change.

I don't think so. I would think that this is precisely the circumstance in which the positive feedback mechanism of "runaway" sexual selection can be amplified, in which a minute and functionless bias in female preference (which can be presumed to vary in the population) can swiftly lead to morphological change in SMRSs, rather than stabilizing selection. What is crucial, if breeding is to occur, is not that SMRSs stay the same, but that SMRSs track each other in male and female. Once the female bias is amplified to nonnegligible relative frequency in a subpopulation (which is ensured by the bottleneck), there is every adaptationist reason for the genes for the male response to that bias to track that bias—even at the cost of "cutting itself off from the fullest spectrum of potential mates," a cost invisible to myopic local selection in any case. Sexual selection is a real and important phenomenon. It is triumphantly part of, not a challenge to, adaptationism—as highlighted in the role of the peacock in Helena Cronin's (1991) title. So I'm delighted with Crowe's "just so story" about guineafowls, which I see as exemplifying the sort of good adaptationism I defend.

Gradualism

Now to gradualism, (c): As I noted in DDI, gradualism must be distinguished from what Dawkins calls "constant-speedism." One can maintain that evolution sometimes runs very fast, and sometimes very slow, but still is gradual, even when it is running very fast. The denial of *this* doctrine is one version or another of saltationism, positing leaps rather than steps, and here one must tread cautiously. Everybody agrees that several different kinds of isolated leaps are

possible: A small, indeed, single step in mutation space can be a rather large step in phenotype space, for instance, adding a whole, well-formed digit, or changing the color of the whole organism, or—to take the truly monstrous cases—building a leg where an eye should be, or creating an extra head. It is even possible, indeed frequent, that single, small mutations can virtually ensure speciation, by disrupting the SMRSs that are, as Paterson vividly puts it, the glue that binds species together. Another way species can become distinct in the absence of the imposition of geographical barriers is a slight change in behavior. If some members slightly prefer (for no good reason at all) to mate at night, and others during the day, this assortative mating can swiftly lead to reproductively isolated subpopulations living together in the same region. Such events create quite abrupt macroscopic changes in the selective environment, and hence in the organisms that thrive in them, but do not count as saltations of the sort denied by gradualists. So, as Crowe says, speciation can be an incidental, accidental effect of adaptation. This does not go counter to anything any adaptationist wants to maintain, so far as I can see, but readily becomes one sort of event that typically occurs in the cascade of small steps that mount up to make for eventual large differences—like the differences between a cow and a whale. Speciation is not itself an adaptation, but it creates an important ratchet or one-way-valve, preventing the random walks of recombination and mutation from returning to the point of departure, and hence forcing further developments to wander down distinct paths. So far, Crowe and I are in complete agreement.

I appreciate his diplomatic attempt at a rapprochement with Gould and his camp, encouraging me to "admit the secondary importance of possible rapid nonselective and/or macromutational change, especially coincident with speciation." But I already did so. As I noted in DDI (294–298), the creation of speciation bottlenecks is indeed a major ratchet in evolution, but it is the bottleneck, not the speciation, that does the work. I also note that a single mutation-event can be a transposition or doubling or other single-step event (287), rather than a single base mutation; and that molecular evolution in unexpressed DNA can accumulate gradually over time in an undirected way, but then come to be expressed, with dramatic (but

usually fatal) results (288). In short, in DDI, I think I acknowledged and even highlighted all the points that Crowe urges me now to acknowledge. And since Gould himself has been most insistent that he never intended his view to be read as a defense even of macromutation ("Punctuated equilibrium is not a theory of macromutation," 1982, 340, quoted in DDI, 289), I think he would cringe at Crowe's claim that he has defended the view that "evolutionary change, especially during speciation, can occur in large, effectively saltatory, steps." In his *New York Review of Books* attack on me, Gould's greatest wrath (which is saying a lot) was concentrated on my claim that he had once flown a saltationist trial balloon. My friendly advice to Crowe: don't use the word "saltatory" within earshot of Professor Gould.

Substrate-neutrality

What about (d), substrate-neutrality? Crowe takes up this feature in his discussion of my third aim, my defense of memes as Darwinian analogues in cultural evolution of genes in biological evolution. He asserts that natural selection cannot be a substrate-neutral process, but since the grounds he gives for this don't seem to me to be relevant, I surmise that once again we have a failure of communication on my part. Of course any particular instantiation of an algorithmic process is not substrate-neutral—you can't run Mac software on a PC, for instance—but the power of the algorithm (bubble sorting, or word-processing, or morphing or whatever) is independent of the substrate of any particular instantiation. Crowe is not the only biologist who has arrived at this misreading of my point. Orr's similar claims (1996) also land wide of the mark, as I have explained (Dennett 1996).

Crowe's discussion of cultural evolution is so compressed—a single paragraph incorporating many claims with scant support—that any useful response from me would first have to extrapolate boldly from his few remarks, and then (guessing that I had captured his intent) compose rebuttals to these views of my own devising. Not likely to be convincing to anybody. So let me do the next best thing by including a generic response (from Dennett, unpublished) to

what may or may not be two of Crowe's points. (*Seager* also has doubts about memes, which I will discuss briefly in the context of my other comments on his essay.)

One often hears it said that the ways in which cultural entities evolve are profoundly un-Darwinian. Two claims, in particular, are often presented as if they carried the day: Cultural evolution, unlike Darwinian evolution, is "Lamarckian," and cultural evolution, unlike Darwinian evolution, is replete with "horizontal transmission"—that is to say, design elements can hop freely from lineage to lineage, not bound by the requirements of heredity. Once reptiles and mammals have gone their separate ways, reptile innovations cannot jump to mammals, but only to descendant reptiles, but this restriction does not exist in cultural evolution. I have sometimes wondered why we don't hear more about a third disanalogy: cultural ideas don't reproduce sexually—mama and papa ideas getting it on to make little baby ideas of both genders. Probably we don't hear it because it would wear its disingenuousness on its sleeve—a lazy (or desperate) stab at something that would excuse one from having to think further about the prospects of a Darwinian account of culture. *Sexual* reproduction is not, after all, an obligatory element of Darwinian evolution; surely 99% of all the Darwinian evolution that has ever occurred on this planet was among asexually reproducing replicators, and however large sexuality looms now, it is itself an evolved feature, not a precondition for Darwinian evolution. So the absence of sexual reproduction in the memosphere is no challenge to neo-Darwinian explanation. But exactly the same point can be made about the purported disanalogies of Lamarckianism and horizontal transmission or anastomosis (lineage joining).

Let's consider Lamarckianism first. Neo-Darwinian orthodoxy, since Weissman, declares that characteristics acquired through use cannot be transmitted genetically to one's progeny. Darwin himself, notoriously, was quite happy to countenance this feature of Lamarckianism, but he has long been deemed in error. Weissman's distinction between germ line—roughly, eggs and sperm—and somatic line cells—all the rest—has proven itself over and over, and the doctrine that there are no avenues by which somatic line innovations could enter the germ line is indeed a textbook verity, although

various exotic possibilities have been seriously discussed in the literature, and arguably exist in some restricted quarters. But notice that this, the orthodox, way of identifying Lamarckian phenomena (as things that don't happen) applies crisply only to multicellular organisms. What counts as a Lamarckian phenomenon in the world of bacteria, archaea, or in the world of viruses? In the case of a virus, which I have described as just a string of DNA with attitude, the line between soma and germ line is nonexistent. Something that changes the structure of an individual virus string can be called a genotypic change—a mutation—if it is passed on in replication, and otherwise a mere phenotypic change. It is not that such a line can't be drawn, but it becomes a line that prohibits nothing. The claim that Lamarckianism has been vindicated in the world of viral evolution would thus be Pickwickian. And since memes are no more multicellular than they are sexual, the fact that there is no clear way—no "principled" way, as they used to say at MIT—of distinguishing mutations from phenotypic acquisitions hardly shows that they are disqualified from a neo-Darwinian treatment. Most—much more than 99%—of the life forms on this planet have evolved under just such a regime, and neo-Darwinism certainly covers their evolution handily.

The same verdict applies to anastomosis, although this is a recent and ill-appreciated discovery: There is lots of horizontal transmission in protist and bacterial evolution—a fact that plays hob with attempts to define separate bacterial lineages in a "principled" way—and once again, the bulk of the evolution on the planet has been among just such tiny bits. Once we shift our focus away from our own multicellular, sexually reproducing lineages to the more numerous lineages on the planet, these standard objections lose much if not all their force. Memes are indeed not very much like elephant genomes, but so what?

Paul Dumouchel wonders if all Good Tricks are Forced Moves. I have one perhaps trivial difficulty with this very interesting paper, which I must expose at the outset. Whether or not Dumouchel's understanding of a Good Trick and a Forced Move is itself a good trick, it is not quite my trick. I meant slightly different things by these terms, and I found myself going cross-eyed trying to do the requisite

translations when reading his essay. So I may not have understood him, but rather some artifact of my own reconstruction. Briefly, I claim that a Good Trick has a wide basin of attraction (or Mt. Fuji of attraction) in Design Space, so that many paths lead to it, from many different starting points. A Forced Move is when your options are reduced to a few; all but one of which is fatal. It is the best move, trivially, but hardly the sort of move one would want to have to choose on a regular basis. As the Godfather would say, you have been made an offer you couldn't refuse. That reverberating phrase, by the way, nicely captures one of the ideas at the heart of Dumouchel's paper. Might it not be true that the more you know about the world, the more *all* your choices come to be seen to be Forced Moves, since the apparent alternatives are either not really available at all, or turn out, on further examination, to be dominated by a Forced Move after all? Consider the home of the Forced Move, the chessboard, and recognize that against perfect opposition, you may be lucky to have *any* move that isn't part of some supersized mating net. In the unfeasible but imaginable algorithm for playing perfect chess, you look at the total decision tree of the game and, if you are white, see if any of the sixteen available first moves is colored white (meaning that there is an unbeatable path to victory by staying on the white moves, one of which will always be available). If there is not, then you must choose a gray move (stalemate guaranteed)—any one you like, if any are available. If there is no gray move, you must choose a black move and pray that at some time your opponent slips up and lets you find a gray or white move. Any such move is a forced move, of course, since it is the only escape from the mating net that began when you agreed to play chess. But for Dumouchel, Forced Moves are forced "because they are all there is."

Stuart Kaufmann has toyed with the idea that all evolution is a set of Forced Moves (in Dumouchel's sense), but it is Brian Goodwin, his sometime Santa Fe colleague, who is the exponent of the view that there really isn't any—or much—optionality in evolution; there only appears to be. The *apparently* available alternatives are not really available at all. Actualism threatens, Dumouchel notes, but so does an evaporation of our explanation (or is it only the illusion of an explanation?) of biological design. Selecting the best out of a field

with only one candidate is no different from selecting the worst, so it can't explain why there seem to be such *excellent* designs in the world. If so, then natural selection could not explain the good design in nature, since it would all be just a charade that it had grown out of a genuine exploratory process of R&D. It is like the jungle boat ride in Disneyland, running on underwater railway tracks and presenting us with phony choice point after phony choice point. The destination, and all the narrow escapes, have been foreplanned and designed as a single path.

Dumouchel notes that this view—Goodwin's or maybe Kauffman's—parallels one of our old friends from the free will literature: "could *evolution* have done otherwise?" Observe, with Dumouchel, that the issue is divorced from classical determinism, both here and in the free will debate (or at least so I have claimed). The alternative to Goodwin's frankly mysterian view is not that evolution proceeds by *genuinely indeterministic* exploration of Design Space, but just that it proceeds by a quite probably deterministic but nonetheless pseudo-random or chaotic canvassing of actual possibilities. (Jacques Monod made the mistake of supposing that unless mutation in natural selection was genuinely random—quantum random, you might say—the theory could not explain the evolution of design. Do we know that this is a mistake? I submit that we do, and that a host of successful demonstrations of evolution by artificial life programs reinforce the theoretical conviction that pseudo-random generation of diversity is quite sufficient.) As Dumouchel says, it is probability that comes to the rescue to salvage the needed distinction between fitness and survival. Many with good designs succumb and many with bad designs flourish, but *over the long haul* and *over the whole population* the good designs will rise like cream to the top, where they can be retrospectively crowned, as the good ideas we may or may not have thought them to be.

I am pleased that Dumouchel has taken this opportunity to introduce Polya's Urn into philosophical discussion. It is a multi-use example of considerable power. Not only is it true, as he says, that the eventual value of the ratio of white to black balls in the urn is a function of nothing but the history of sampling, but it is also true that there is no telltale feature of the sampling histories themselves

that could be used to distinguish Polya's urn from a standard urn, which begins with a determinate number of white and black balls and is sampled repeatedly (without doubling the sample). But I think Dumouchel slightly misspeaks when he says "once a particular value has been 'chosen,' it will always come to fixation after a period of fluctuations." It is rather that as the population in the urn grows, the effect of any further drawing has a diminishing effect on the overall proportion. But at no time is a particular value "chosen." The value always evolves, just slower and slower and slower. So I like the example but I am not quite sure what point he wants to make with it.

I have recently concocted a new example of my own; drawing once again on my favorite topic of chess-playing computers, to illustrate a point this is at least close to Dumouchel's. Install two different chess-playing programs on your computer, and yoke them together with a little supervisory program that pits them against each other, game after game, in a giant tournament, and let some arbitrary function—readily available from the pseudo-random number generator bundled with the computer—adjust the starting state of each program at the outset of each match. (I'm not supposing these programs are learning programs—I just want to get "rewinding the tape of life" into the picture so these two programs don't just play the same game over and over and over again. I want them to play a huge variety of games from their implied competencies.) Now sit back and look for patterns. Perhaps A always beats B; perhaps not. I don't care for the moment about that pattern. I want to focus on the standard patterns of chess strategy, on such facts as the near inevitability of B's loss in any game where B falls a rook behind, as the fact that when A's time is running out, A searches less deeply in the remaining nodes of the game tree than it does when in the same local position with more time remaining, and so forth. There is a cornucopia of regularities of this sort to detect, and they have the effect of highlighting moments in the unfolding of this deterministic pageant that otherwise would be all the same. After all, the two programs yoked together in "combat" under conditions determined by the next few digits of the pseudo-random number generator form a single utterly deterministic automaton unfolding in the only way it can, with no

"real" forks or branches in its future; all the "choices" made by A and B are already determined.

If you look down in the computer's engine room, the CPU, all you find are fetch-execute cycles, and they're all alike; each one causes the next. At a higher level or perspective, you get chess moves—white's failure to detect the sacrifice-opportunity at move 5 caused white to terminate search earlier than otherwise, which in turn caused black to have less time to ponder without the clock running, which caused black to miss the deep continuation. . . . That led black to make the dumb move that caused white to win. So white's earlier failure led "inexorably" to white's eventual victory. It would happen again and again—*if* white and black were placed in *exactly* the same state. Of course white's initial "failure" to detect the sacrifice-opportunity was in turn caused by each earlier state of the system, and so forth. It is only when we gather data from lots of runs—when we "wiggle the events" as David Lewis has put it—that we can see the higher-level patterns that in one sense justify our conviction that not all fetch-execute cycles are created equal. Some of them are "pivotal." In what sense? In the sense that they, and not others, appear again and again in varied reruns of the tape of life. There are patterns that emerge from multiple runs that tell us something about what is "important."

As Hume showed, we cannot rationally defend our assumption that the future will be like the past. It is interesting, nevertheless, that Mother Nature, or evolution, proceeds *as if* it was assuming that the future will be like the past. It cannot help doing this, of course, just as we cannot, as Hume notes. It is not rationally defensible, but there is something to be said for it, if only in the form of a rhetorical question. Recall the proposed campaign slogan when the notoriously dim-witted Gerald Ford was running for president. "If I'm so stupid, how come I'm President?"

Dumouchel and I are on the same trajectory so far, I think, but then I lose him when he says: "If Good tricks and Forced Moves can be pried apart, it would in principle be possible to show that it is because certain moves are good that they were selected, rather than conclude that it is because they were selected that they are good." I am not sure I understand what Dumouchel calls his "naturalized

Kantianism''; he says both that we need to answer the question—Are all Good Tricks Forced Moves?—and that we cannot. Moreover, he says that this fact ''indirectly supports the claim that our intentionality and reason are products of natural selection.'' I'm not sure why. Dumouchel's claims about the implications of this for evolutionary ethics are a topic that I am hoping to survey more carefully in the near future, but it is too complicated to address here. (I offer a few opening suggestions in my comments on *Mooney* below.)

Evolution and Intentionality: Millikan, Kenyon, Seager

Ruth Millikan's essay addresses both topics, and hence serves as the perfect segue, especially since her views and mine on these topics have been so closely intertwined over the years. She wonders whether the residual differences between our views on intentionality and indeterminacy can be made to evaporate, and toward that end she identifies several apparent points of contention. (1) I take the intentional stance to be ''more basic'' whereas she takes the design stance to be more basic. (2) I recommend adopting the intentional stance toward natural selection itself (Mother Nature) and she finds this otiose. Finally, (3), I think indeterminacy is a ''holist'' and presumably global phenomenon, whereas she thinks it is local, and not such a big deal. I do think we can close the gap, thanks to her efforts and those of Kenyon (and the others, in the discussions in Newfoundland).

First, I agree with her that the design stance is more basic, in the sense she defends. Not all animals display ''real rationality patterns,'' she claims; tortoises, for instance. I agree that there is a huge difference in the versatility and richness of the perceptuo-behavioral talents of different species, and I wouldn't be all that surprised if there were some theoretically meaningful way of identifying a subset of them as those whose rationality patterns deserved to be called real, but I haven't yet seen such a scheme that can persuade me that my more open-ended way of identifying intentional systems (which includes not just the tortoise but the thermostat, after all) is a tactical error. Millikan doesn't tell us what convinces *her* that a species is one of those designed to have real rationality patterns, and there are in

fact lots of vexing penumbral cases in between the tortoises and us—supposing with her that tortoises really do fall on the low side of some important divide. Filling in the gaps here is a philosophically significant chore, by the way, as Seager's essay makes clear. If we are the only species on the happy side of the "real rationality patterns" divide, we will have to make sure that our evolutionary story doesn't stop short of reaching *us*. (Curiously enough, in these days at the turn of the millennium, there are still philosophers and even scientists who harbor hunches about how the gap between *Homo sapiens* and all other animals needs a skyhook to be closed.) Millikan is right in any case that it is no accident that the entities that succumb to the intentional stance *projectibly* must have been designed to do so. "From enough apparently rational behavior one can infer design for rationality, just as one can infer design for seeing from good sight. . . . An intentional system *is* a designed system." Since this is what she means by saying that the design stance is more basic, and since I accept it, I can go along with her gladly on this point, but doing this then obliges me to confront (2): Since the process of natural selection is not itself a designed thing (though it is a designing thing), it cannot be an intentional system, can it? Yes it can, because it meets (or evades) the design requirement in a unique way.

Use of the intentional stance in biology—the Mother Nature stance, you might say—is at least a convenient compactor of messy (and largely unknown) details into a useful interpretation-label. It is *as if* Mother Nature had this or that "in mind"—and Millikan herself lapses into just this usage at one point. Free-floating rationales abound in biology. Evolutionary game theory exploits them in a big way. When an organism's environment is largely unpredictable (to whom?—to the process of natural selection, really), evolution wisely installs learning mechanisms instead of rigid tropisms, etc. But Millikan asserts with as much plausibility as emphasis: "*There is nothing in Nature analogous to beliefs and nothing that so much as reminds one of inference.*" What gives? Surely Millikan is right that the process of natural selection is not a designed, structured, representing system like a brain, but all the same, I think she is wrong, in a very interesting way, when she claims that "nothing so much as reminds one

of inference." As Sherlock Holmes, the patron saint of inference, famously said, once you have eliminated all other possibilities, the one that remains, however improbable, must be the truth. Is that not an inference? Does not Mother Nature eliminate all other possibilities, on a vast (not actually Vast) scale, thereby "inferring" the best design? When Deep Blue eliminates a few billion legal moves and comes to rest on one brilliant continuation, it surely reminds Kasparov of inference! Natural selection, like a clever stage magician, hides almost all her trials, giving rise to the illusion of a miracle. She accomplishes by brute force (plus the utterly indispensable ratchet of selective accumulation of local progress) what otherwise would require foresight and intuition and brilliance. Deep Blue was designed to eliminate huge haystacks, thereby finding needles. Natural selection was not originally designed; it was a fortuitously emergent process of replicator-sorting. But, unlike Deep Blue, it happens to be a self-improving, self-redesigning phenomenon: It has made crane after crane after crane, becoming ever more efficient, even to the point of creating foresight—in us—so that it has bootstrapped *thoughtful* design into existence. Deep Blue is an intentional system—even for Millikan, if I understand her aright. Compared to Deep Blue, what natural selection lacks in pedigree—it is not *from the outset* the product of an R&D history of intentional design—it makes up for in forward-looking versatility and the capacity to "learn": It is getting "smarter and smarter" in its subparts.

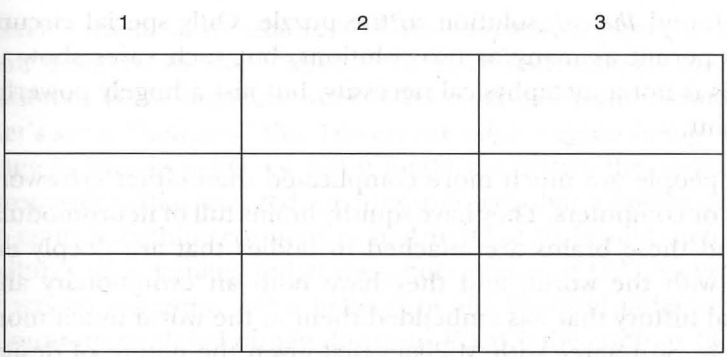
Millikan also challenges what I have said about the varieties of indeterminacy that can arise in those histories of natural selection that parallels the history of R&D by engineers. She sees these indeterminacies as "temporary and uninteresting," two flexible and perspectival adjectives just begging to be the ground of a rapprochement. I myself find these indeterminacies, which she describes well, deeply interesting, precisely because all the "accidental junk thrown up by mutation" that comes along for the ride doesn't just sully the purity of the "definite" way that the animal has been designed to operate; it is the indispensable seed bed out of which improvement (or just revision) can emerge (DDI, 407–408). Fortunately, it is ubiquitous (and so, like taxes, both temporary and permanent

at the same time). Since Millikan notes that "what counts as the correct causal explanation [of what the frog's eye tells the frog's brain] is vague in a way that probably cannot be eliminated in any principled way," I'm hoping to get her now to agree with me that these indeterminacies are not quite so temporary or uninteresting as she has made out. So, having clarified my grounds, thanks to Millikan's good challenge, I remain unrepentant in my usage here, though I do pay a price: witness *Crowe's* all too widely shared interpretation of my engineering claim.

That leaves her claim that they are "local, not holistic," and this brings us to (3), which I will discuss in the context of belief-attribution, not evolution. We are actually close to agreement here, too, I am happy to say, since she agrees with me that the "strong realist" should *not* claim that there is a fact of the matter in those cases where there is, by her lights as well as mine, indeterminacy. The question is: How much indeterminacy is there, really, and what is its importance? *Kenyon* expresses parallel doubts, differently cast, about my line on indeterminacy, and I think my Quinean crossword puzzle may be just the tool needed here to jimmy us all into agreement.

Is 1 down, the retentive membrane, "web" or "sac"? Aha! The puzzle as a whole has two solutions. (It is not, I grant, a thing of beauty; but devising a puzzle with two solutions that are "tied for first place" took some serious fussing on my part. It was an instructive exercise; anybody who thinks that indeterminacy of radical translation is bound to be a common and stable phenomenon in the real world will come to see how remarkably powerful the constraints of even such a simple task as this are.) Whatever its aesthetic shortcomings—and I invite all readers to rise to the challenge and send me something more elegant to use—it instantiates a few key claims.

1. Anybody who asks, "Which word is 1 down *really?*" stands convicted of a certain sort of overreaching realism. *There is no fact of the matter.* I deliberately set it up so there wouldn't be a fact of the matter. For instance, I didn't compose the puzzle with one set of answers (the historically first or original answers and "hence" the *real* answers) and then cast about for another set. I worked out the two



Across

1. Suck the resources out of
2. Epoch
3. Sleep furniture

Down

1. Retentive membrane
2. Earlier
3. For some kids, a best friend

Figure 15.1
Dennett's Quinean Crossword Puzzle

solutions together, drawing from a list of pairs of short words I'd drawn up that had similar meanings.

2. The reason it is possible to construct such a puzzle is that there are norms for definitions that admit some flexibility. Both solutions include words that just barely fit their definitions, but the conspiracy of the surrounding fit (the holism, to give away the punch line) pulls the words into two quite stable configurations. And I daresay everybody will agree that there is not going to be a third solution that competes with either of these. In general, the cryptographer's constraint holds: If you can find *one* solution to a puzzle,

you've found *the only* solution to the puzzle. Only special circumstances permit as many as two solutions, but such cases show us that this is not a metaphysical necessity, but just a hugely powerful constraint.

Now people are much more complicated than either crossword puzzles or computers. They have squishy brains full of neuromodulators, and these brains are attached to bodies that are deeply entwined with the world, and they have both an evolutionary and personal history that has embedded them in the world much more invasively. So I agree with Millikan that given the nature of design constraints, it is unlikely in the extreme that there could be different ways of skinning the cat that left two radically different, globally indeterminate, tied-for-first-place interpretations. Indeterminacy of radical translation is truly negligible in practice. Still, the principle survives. The reason we don't have indeterminacy of radical translation is *not* because, as a matter of metaphysical fact, there are real meanings in there (Quine's "museum myth"), but because the cryptographer's constraint just makes it a vanishingly small worry. When indeterminacy threatens in the real world, it is always just more "behavioral" or "dispositional" facts—more of the same—that save the day for a determinate reading, not some mysterian "causal power" or "intrinsic semanticity." Intentional interpretation almost always asymptotes in the limit at a *single* interpretation, but in the imaginable catastrophic case in which dual interpretations survived all tests, there would be no *deeper* facts to settle which was "right." Facts do settle interpretations, but it is always "shallow" facts that do the job. That is all that Quine and Davidson and I have ever wanted to hold out for, I think. It is all that I have wanted, in any case. And so far as I can see, this brings me right alongside Millikan; we agree that design features, which are the crucial "shallow" constraints, leave no room for doubt in general.

I remain skeptical, however, about Millikan's optimism about how the indeterminacy always stays local. The rationality generalization is of course both the joy and the bane of AI and expert systems researchers. When people confront expert systems, they naturally, ineluctably, infer more rationality than is there. Consider Douglas

Lenat's (Lenat and Guha 1990) CYC—a system designed to be rational, designed to have "beliefs" that are both globally and locally rational. What happens when CYC exhibits "local" irrationality? Let's say it "believes" that lawyers are adult human beings that engage in various activities, but it might not notice the anomaly of a proposition that asserted, or presupposed, that a lawyer appeared in court in bathing costume, or had an IQ of 50. Would such "local" failures of rationality indict the whole system of CYC, showing that it was an impostor, not a believer at all? Do local indeterminacies mount up to global disqualification? What *is* the content (if any) of CYC's lawyer-data-structures? Are these cases in which the system's mechanisms are well designed, but "laboring under conditions that fail to support them properly"? I am not convinced that Millikan can make this distinction, on which her case depends. I'm also not convinced that she can't. (I have recently come to think that Dan Sperber, "Apparently Irrational Beliefs," 1985, offers a wealth of novel insights on this family of issues. See especially his suggestions about "semi-propositional representations.")

Before leaving Millikan's rich paper, I want to applaud her point about how utterly unlikely an *exaptation* for rationality would be (cf. *Seager's* discussion of Chomsky). I also want to agree with her note 5, that there is no one perfect ideal of rationality, and to acknowledge as well her excellent examples of William and James, whose belief-relevant dispositions do not have a pure intentional explanation. Add these examples to the case of a belief causing someone to blush (see my discussion of *Viger*) and the case of the computer that thinks it should get its queen out early and we have a panoply of different phenomena filling in the penumbral territory between the purest intentional stance and a design stance that uses intentional characterizations simply as labels for the design elements under discussion. Which of these have "legitimate intentional explanations"? I see a gradualistic ramp of cases, but am open to arguments showing that there are important thresholds to be marked.

Tim Kenyon points out that Chomsky's challenge to Quine—"why suppose that the under-determination of content ascription is more exotic than that found everywhere?"—was never answered (but just

rebuffed) by Quine, and suggests that I might be invited to make good on my mentor's promissory note. My Quinean crossword puzzle is the downpayment; here's the rest.

Way back in *Brainstorms,* I developed an example that highlights the difference between physical stance and design (or intentional) stance disagreements:

> Suppose Jones and Smith come across a particular bit of machinery churning away on a paper tape. They both study the machine, they each compile a history of its activity, they take it apart and put it back together again, and arrive at their pronouncements. What sorts of disagreements might there be between Jones and Smith?
>
> First we might find them disagreeing only on the interpretation of the input and output symbols, and hence on the purpose or function of the Turing machine, so that, for instance, Jones treats the symbol-features as numbers (base two or base ten or what have you) and then "discovers" that he can characterize the Turing machine as determining the prime factors of the input numbers, while Smith interprets the symbol features as the terms and operators of some language, and has the Turing machine proving theorems using the input to generate candidates for proof sequences. This would not be a disagreement over which Turing machine had been realized, for this is a purely semantic disagreement; a Turing machine specification is in terms of syntactic relationships and functions only, and ex hypothesi Jones and Smith agree on which features are symbols and on the rules governing the production of the output strings. In principle a particular Turing machine could thus serve many purposes, depending on how its users chose to interpret the symbols.
>
> More interesting and radical disagreements are also possible, however. Jones may announce that his device is TM_j, that its input and output are expressions of binary arithmetic, and that its function is to extract square roots. However, let us suppose, he proves mathematically (that is, on the basis of the machine table he assigns it and not the details of the engineering) that the program is faulty, giving good answers for inputs less than a hundred but failing periodically for larger numbers. He adds that the engineering is not all that sound either, since if you tip the machine on its side the tape reader often misreads the punched holes. Smith disagrees. He says the thing is TM_s, designed to detect certain sorts of symmetries in the input sequences of holes, and whose output can be read (in a variation of Morse code) as a finite vocabulary of English words describing these symmetries. He goes on to say that tipping the machine on its side amounts to a shift in input, to which the machine responds quite properly by adjusting its state-switching function. The only defect he sees is that there is

one cog in the works that is supposed to be bent at right angles and is not; this causes the machine to miscompute in certain states, with the result that certain symmetries are misdescribed. Here there is disagreement not only about the purpose of the machine, or the semantics of the language it uses, but also about the syntax and alphabet. . . . The two may still agree on the nature of the mechanism, however, although they disagree on what in the mechanism is deliberate design and what is sloppiness. That is, given a description of the physical state of the machine and the environment, and a physical description of the tape to be fed in, they will give the same prediction of its subsequent *motions,* but they will disagree on which features of this biography are to be called malfunctions, and on which parts of the machine's emissions count as symbols. ("The Abilities of Men and Machines," 258–259).

A saying in the world of software development expresses the crux of such disagreements rather well: "it's not a bug, it's a feature." The reliance on one or another ideal of *excellence* (or *rationality* in the case of cognitive excellence) opens up the space for further indeterminacy once we move above the physical stance. (I *think* this is what Kenyon means, in the end, by his proposed principle: "The truth conditions for content ascription cannot be purely physical, on pain of its being possible that there are psychological truths that are unverifiable in principle." An unremarked difference between the physical stance and the design stance is that no oxygen atom is "defective" or subpar in any way. They are all "perfect." If they weren't, there would have to be a physical design stance of sorts, and it would be vulnerable to all the caveats and fallings-short that arise in the world of design. In fact, even design stance attributions go through quite determinately as long as design elements are sufficiently robust, sufficiently close to their "specs." In practice no problem arises about how good a resistor or capacitor has to be to count as a resistor or capacitor, and designs composed of such elements don't have to be massively redundant in order to cope with the inevitability of ubiquitous imperfection. If we were not products of natural selection, if we were all Model Z3J Electronic Believers, doing our thing (not things) in a standard environment (so that the prospect of evaluating CYC in *our* multifariously convoluted human environment didn't arise), hermeneutics would be child's play. As it is, hermeneutics is a haven for indeterminacies of the sort that

thrive on the inevitable fact that one programmer's bug is another programmer's feature.

Kenyon is dubious about my efforts to get philosophers to lighten up about such traditional issues as ontology and truth:

> Those uneasy with the suggestion that our theory of psychological explanation implicitly incorporates a theory of voices or a theory of salt will react sharply to Ryle's and Dennett's maneuver. The obvious objection is that we already have a word that means "true with a grain of salt" or "true in one logical tone of voice." That word is "false."

Yes, that is the way philosophers are apt to think. And Kenyon is right that neither Ryle nor I go to the lengths we would be obliged to go to satisfy those uneasy folks. But we do what we can to add to their unease, by showing them that the very verdict they are so tempted to declare about our cases will also have to be rendered for less controversial endeavors they might regret abandoning. Their bracing allegiance to the true-false dichotomy is not quite so comfortable to live with as one might think, and Kenyon is helping to see what the issues are by holding the tradition he was trained in at arm's length, throughout his essay.

> In contemporary philosophy of mind, emphasizing the central epistemic role of behavior is rather like admitting to being a liberal in American politics: You must immediately backtrack, qualify your position, and distance yourself from everyone else who ever said anything similar. But the idea is really quite benign.

So true. I also endorse Kenyon's argument about color and subvening bases (*modulo* some sophistications that Kathleen Akins would rightly insist on. See also my comments on *Lloyd,* below). I'm glad to have a more patient defense of this than I have undertaken. I also applaud Kenyon's lengthy note (don't miss it) on Fodor on Stich's Mrs. T. He has seen exactly what is untenable in Fodor's response. And his comments on Crispin Wright and Michael Dummett persuade me that I have been underestimating a literature that has more to show me than I had thought. The one point where I feel the need to issue a (minor) correction is Kenyon's opening overstatement of my view about misrepresentation (by the frog's eye): I do not *maintain* that "misrepresentation can always be dissolved

in this [disjunctive] manner"; I am issuing a challenge to those who find this claim a reductio: Figure out, then, how to block that deflation-by-disjunction.

William Seager's essay covers so much ground, running from evolution to intentionality to consciousness to ontology, that I have had trouble finding a good place for it in the marching order. I find so much to admire in his analysis of the issues that I feel churlish refusing to go the last step, abandoning the Scientific Picture of the World for Surface Metaphysics. I do have my reasons, though, and Seager comes within a whisper of articulating them. But first, let me show how far Seager leads me before I resist.

Consider his master stroke: his contrast between the demands of naturalization in chemistry and psychology. In the case of chemistry, you don't *have* to have a prior understanding of the higher level (the chemistry-level concept of valence) in order to understand, from the bottom up, as it were, the lower level: "someone could (albeit inefficiently) learn what valence is without bothering about developing a prior acquaintance with chemistry—there is no need to understand chemistry in order to understand the physics of valence." But *some* prior understanding is required, however; in such a case, he notes that "no one could understand the physical account of valence without already understanding what explanation is supposed to be." This is no threat to naturalization of chemistry since the mentalistic or intentional concept of explanation is not itself a chemical notion. But when we try to make the same move in psychology we unavoidably step on our own toes:

> You can't understand what a mind is unless you already know what a mind is, since you can't understand mentality without understanding the intentional stance, which requires you to already understand a host of essentially mentalistic concepts.

Is this really a problem? It certainly seems to be a problem, and this is a fine way of expressing a background fear that has haunted many a move in the philosophy of mind. Now that it is so well exposed, we can confront it. (Another version of this fear is the unease people often feel with my account of heterophenomenology. It seems to some of them [but not to *Thompson,* see below] that I *must* be doing

something viciously circular in giving pride of place to the third-person perspective over the first-person perspective.) Seager's is a better setting in which to consider the issue, I think, since it abstracts away from the overheated topics of qualia and zombies and presents itself as a minimalist demand on *explanation:* If a purported naturalizing *explanation* of minds as phenomena in the world presupposes that the *explainee* (the one to whom the explanation is supposed to be illuminating) already understands mentality, this will be an unacceptable surrender to what Seager aptly calls Plato's problem. And as he notes, this shows us how to make good use of Davidson's claim that "discovering the neural 'correlates' of mental states does not *explain* the physicality of the mind."

Seager sees, correctly, that I have tried to articulate an alternative naturalism that "must forsake rule bound naturalization." (This is a point upon which there seems to be a convergence of agreement, from *Ross, Lloyd,* and *Viger,* at least. I find this heartening.) And it proceeds, as he says, with the help of Darwin. Exactly right. Nobody has articulated better than Seager the underlying rationale of my Darwinism, especially in this context. Let me underline what he has said to make sure we secure this key point once and for all.

In a world without perspectives, without minds, without errors, a process of replication and competition initiates itself. This process inevitably yields lineages of *(proto-)wanters*—competitors for the resources needed for replication—that achieve efficiency by evolving sense organs that provide them with *(proto-)beliefs* about those resources and the means to secure them. If you're going to move, you're going to need reconnaissance, and so informavores, pattern-detectors, hemi-semi-demi-minds are, as Seager puts it, "a fat evolutionary target." For the reasons he enunciates, there is a large—maximally large, I would say—basin of attraction for this Good Trick.

But Seager sees problems with my project of extending this explanatory account all the way up to us. There is a big gap to be traversed—including, perhaps, the notorious "explanatory gap" of consciousness itself—between, on the one hand, the proto-minds of bacteria and other thermostat-like strivers for reproduction, and, on the other hand, our indefinitely complex and versatile minds. If

this gap is filled with mere "accident or spandrel" we are right back to a mysterian acquiescence in brute fact. (Seager is deliciously on target when he points out that Flanagan is indeed a mysterian *malgré lui* when he blandly announces that "some patterns of neural activity result in phenomenological experience; other patterns do not.") And that explains why I have devoted so much attention to filling that gap with solid adaptationist extensions of the Darwinian reasoning that gets us rather uncontroversially to the simplest kinds of minds.

Seager thinks, however, that I fail to bridge this gap, even sketchily. He sees that there is *hope* for a "quasi-naturalist" filling of this gap by evolutionary psychology, but thinks the task is largely undone. "It leaves behind an outstanding debt to the scientific worldview, which can be only partially repaid by an evolutionary account of the genesis of the behavior patterns that are the targets of the intentional stance." Why only *partially* repaid? Because of the prospect of "accidents or spandrels," apparently. But we don't know that this is a problem. In fact, I think it is not, but let's leave that issue aside, since I want to argue that Seager's "methodological mysterianism, which is general and unavoidable" is not the embarrassment it appears to be in any case. I say there is nothing viciously circular about our inability to explain minds without presupposing an understanding of minds. For that reason, I don't really like calling it mysterianism—though I approve of the way the term dramatizes the issue—since there doesn't seem to me to be anything puzzling left over, anything to wonder about.

The lack of a neutral Archimedean point from which to start is a curiosity, I think, not a puzzle. One way of deflating its importance might be to imitate Descartes's tactic, in *Le Monde,* of describing "a new world" in loving detail and then, in the punch line, revealing that *that* fictional world is our world. They are us.[1] So imagine that we survey the biota of Planet X, elaborating a Darwinian account of the growing sophistication of the minds (or *schminds,* if you insist) of the Xian fauna, thanks to the various arms races of competition among the clever locomotors there, until we have explained—in good bottom-up, no-skyhooks fashion—the eventual emergence of critters on Planet X with language, with open-ended recursive

reflexivity, with cultural sharing of cognitive tools and resources, ... Hmm. They look a lot like us. Don't they? Perhaps what explains *them* explains *us*. Why not? I submit that we can chip away at the limit on inquiry that Seager has identified for us, moving from valence to Planet X zoology, to Planet X psychology, always presupposing *our* intelligence, of course (no use trying to engage a potted palm in scientific inquiry, is there?) but scrupulously restricting our attention to things other than us. To the objection that we are *imposing* our psychology on them, we can blandly reply that if our imposition doesn't work, we will be punished by the falsehood of our predications, and if it does work, we'll have good reason—the best—to conclude that our "imposition" was trivial. Similarly, we may opportunistically "impose" our biology on the (so-called!) fauna and flora of Planet X. If our spade is turned, we will learn something about the limits of our biology; and also about the absence of such limits, if our spade isn't turned.

What limits, if any are there, then, to our simply discovering that we have inadvertently been looking in a mirror? I do not see any, and these considerations suggest to me that methodological mysterianism, once brought into the open, proves to be a bogeyman of no serious concern. (See also my comments on *Thompson* below.) But Seager has performed a very useful service in articulating this subliminal deflector of theory.

Now back to Seager's worry about the only "partial" success of my Darwinian reconstruction of mind. He sees that my account really needs something like memes, and he is as skeptical as *Crowe* is. He gives no more grounds for his doubts than Crowe does, so this is not really the place for a full-fledged defense of memes. It is, however, a place for laying down signposts and promissory notes to just such a defense: Dennett, unpublished a, unpublished b, and various works in progress. (Two international workshops on cultural evolution in 1999, one at Cambridge University in June, and one in Paris in November, should clarify many of these issues.) For the time being, I will provide only one brief rejoinder to Seager's claim that "Memes are products and inhabitants of minds *as such*." I don't think so. The basic phenomenon of memetic transmission and evolution can be built up from a *very* slender base: Mere imitation will

perhaps do the trick (see Blackmore 1999, and Dennett, unpublished b), and although imitation is a more sophisticated behavior than running or digging holes, it isn't rocket science. Quite mindless beings could "in principle" harbor memes—teenagers, for instance. (Joke)

Seager helps us think about the scientific picture of the world and the relations that are deemed to hold there by introducing an excellent thought experiment, the massive bottom-up computer simulation of physics. (This is close kin to my use of Conway's Life World in 1991b and 1995, but has its own additional virtues.) The main conclusion Seager wants us to see drawn from this thought experiment is that "the world has no use for" the higher-level patterns that are visible in the simulation: "the *only* role they have in the world is to help organize the experience of those conscious beings who invent them and then think in terms of them." Not quite, I think, but this way of putting the point focuses our attention on the role of *pattern-detectors* in the world, a topic also central to the essays by *Lloyd,* and *Ross.* The Coriolus *force* is non-existent, Seager notes, but the *pattern* of the Coriolus force is as real as can be. Are the patterns "metaphysically otiose"? What does that mean? *Ross,* after all, offers to define existence itself in terms of such patterns. And what does it take to be a pattern-detector? Consciousness, Seager suggests, but this strikes me as backsliding. "Precisely where the changeover from nonmind to mind occurs is a vexed question (even allowing that the distinction will be fuzzy), but what matters is the point that *patterns* have no role to play in the world unless and until they are taken up in understanding by minds." I disagree. All those simpler, thermostat-like minds are responsive to patterns; they digitize (in the sense of Dretske 1981) part of their interaction with the world, willy-nilly creating encodings of those patterns (whether or not the encodings are *for them*). In Seager's opinion, "Mind cannot be 'just another' pattern." Why not? Perhaps I have missed his point.

Aside from this claim of his on which I have just demurred, I find Seager's analysis of different varieties of emergence right and valuable (see also Holland 1995, 1998 for excellent discussions of what Seager calls "benign emergence"). I also like Seager's whimsical sketch of a "spandrel" vision of consciousness, which supposes it to

be a surd by-product of the blood-cooling machinery of the human brain. He vividly brings out the dilemma that faces Gould: Either he must mean something so attenuated by his spandrel claim that it doesn't do any work—all adaptations were once spandrels (or exaptations—see Preston 1998 and Dennett 1998c)—or he ends up with an utterly preposterous hypothesis. See *Millikan*'s chapter for a similar argument against Gould.

Finally, I don't want to convey the impression that I think Seager's surface metaphysics is out of the question. I take under advisement the recommendation of van Fraassen's constructive empiricism. I get the point, I think, but don't yet feel the itch. My ontological convictions are now in happy disarray, and I simply have no clear sense of how to put together Seager's proposals with the various ideas of *Ross, Viger, Lloyd, Thompson,* and others, but perhaps my reactions along the way will create a pattern that another mind can detect and understand better than I can. I hope so.

Intentionality and Realism: Viger, Ross

Christopher Viger concentrates on the problems I've created for myself by adopting the abstracta-illata distinction I lifted from Reichenbach (see also *Ross*'s chapter, and below). As Viger suggests, people persist in reading as an (extravagant) *ontological* thesis what I took to be a thesis about different explanatory practices in science and their requirements. It is clear that I have underestimated the sources of confusion that Viger usefully sorts out, and I am grateful for his generally lucid and accurate portrayal of my position, but I see one area in which he exaggerates (perhaps for simplicity) my position on the relationships between (folk) psychology and physics: "When distinct explanatory practices are cross-applied," he says, "hybrid monsters result." I don't think all such hybrids are monsters. After all, I'm quite happy to countenance "Her belief that John knew her secret caused her to blush" (1987, 56), and this is a hybrid, part intentional stance (identifying her state via the content of her belief) and part physical (or, arguably, low-level design) stance. The problem arises from . . . thinking that this is a problem! Apparently, many philosophers have convinced themselves that they cannot countenance

such causal claims without going through the Procrustean mills of warring doctrines of supervenience and token identity theories. It is deemed that there is a problem of "overdetermination" along these lines:

> Some complex neural state caused the capillary enlargement that is token-identical to her blushing. Either her belief is *token-identical* to that neural state, or it isn't. If it isn't, then it can't have caused her blushing (that's one cause too many) so it must be *epiphenomenal.* Instrumentalism, by denying the token-identity theory, removes abstracta from the realm of causation, and is scarcely distinguishable from epiphenomenalism.

I have encountered the same blockade in reactions to my thought experiment about the Two Black Boxes (DDI, 412–422). How could it be that the *truth* (or believed-truth, or other intentionally characterized property) of the impulse patterns *causes* the red light to go on when we already have a perfectly complete physical-level account of all the microcausation of each state of the two systems?

But the main point of the example of the Two Black Boxes is to demonstrate the *need* for a concept of causation that is (1) cordial to higher-level causal understanding distinct from an understanding of the microcausal story, and (2) ordinary enough in any case, especially in scientific contexts. With regard to (1), let me reemphasize the key feature of the example: The scientists can explain each and every instance with no residual mystery at all; but there is a generalization of obviously causal import that they are utterly baffled by until they hit upon the right higher-level perspective. (In Seager's fine terms, this is an example of "explanatory emergence"; "complexity does outrun the *explanatory resources* provided by an *understanding* of the simple.") With regard to (2), the contrived example of Two Black Boxes is only artificially clearer than a host of familiar cases having the same logic—uncontroversial cases that philosophers have tended to overlook. An earlier example of mine (in Dahlbom 1993, 216) is the center of gravity of a sailboat, which is manifestly an abstractum and just as manifestly implicated in important *causal* generalizations. I didn't bother to spell it all out before, but since this claim has met with skepticism, I will now do so. What did Connor do overnight to *cause* his boat to be so much faster? He

lowered its center of gravity. Of course he did this by moving gear, or adding lead ingots to the bilge, or replacing the mast with a lighter mast, or something—but what *caused* the boat's improvement was lowering its center of gravity. This is not just casual shorthand; it is the generalization that *explains* why any of these various changes (and a zillion others one could describe) would *likewise* cause an improvement in performance. Differences in the location of a center of gravity cause projectible differences in performance. It is a kind of cause that can be readily isolated by the Millian method of differences, and any philosophical doctrine that denies that this is a good clear case of causation is in trouble. Manifestly, one doesn't need to be a token identity theorist about those centers of gravity to cite them in such contexts. Selection pressure in evolutionary theory, inflation in economics, and a host of other high-level, diffuse (from the mole's-eye perspective of these philosophers) phenomena are perfectly fine causes. That beliefs can cause blushing is just as uncontroversial. Viger is thus somewhat off-target when has me holding the following:

> It is only through the filter of rationality considerations that intentional patterns are visible, and it is for this reason that beliefs and desires have no place in physical explanations.

What is right about it is that it is only via the rationality considerations that one can identify or single out the beliefs and desires, and this forces the theorist to adopt a higher level than the physical level of explanation on its own. This level crossing is not peculiar to the intentional stance. It is the life-blood of science. If a blush can be used as an embarrassment-detector, other effects can be monitored in a lie detector. Pregnancy can be the cause of triggering a positive result in a pregnancy test, and a history of hepatitis can cause telltale effects.

Aside from this point of modification, I find much that is helpful and right in Viger's essay, especially its way of illuminating the fact that ontology, for me, has always been the caboose, not the engine. As Ramberg (1999) puts it, "Such questions, the philosopher's questions of ontology, are for Dennett, as they are for Rorty—and as they were for Dewey, James, Nietzsche and perhaps for Hegel—questions

that get settled after hours, after the real work is done." He suggests that I ought to countenance an "instrumentalist" abstractum becoming (or coming to be recognized as) a "realist" illatum in the fullness of time, or vice versa. The difference I illustrate by contrasting a center of gravity and an atom may be dissoluble on closer inspection. I am tempted. After all, in the history of science, items that began their careers as convenient fictions or instrumentalist abstracta have been promoted (if that is the right word) to the company of illata more than once, and the reverse fate is not unknown, though often with some pushing and shoving and Whig history in the aftermath. Atoms, famously, were but a useful fiction to some of their earliest advocates, and Mendel's genes now have to face banishment to the limbo of instrumentalist idealizations of population geneticists now that more robust (if often less tractable) versions of genes are being manipulated by the microbiologists. Are genes real? This is the Age of Genes. How could they not be real? Well, one of the findings in the Age of Genes is that nothing in nature *quite* fits any of the "classical" definitions of a gene. So should we be eliminativists or instrumentalists or realists about genes? I still think these ontological questions are the *last* questions we need to answer—and their answers will not be very interesting or useful once we've got the science in place, with its various levels of explanation.

Don Ross offers to help me with the metaphysics, and I am tempted to follow his lead. I go round and round on this paper, seeing his points and then watching his points evaporate (for me). In the end, discretion wins over valor. What *ontological* lesson should we draw from my various intuition pumps? I don't know. I *still* don't know. I'm not confident about metaphysical judgments, so why risk taking on baggage I don't deeply understand and am not sure I need? Besides, if Ross is right and there is metaphysical gold in them thar hills, I expect he can mine it and refine it better than I can.

In addition to the welcome support he gives my brusque dismissal of the "merely logically possible" worlds in which various famous thought experiments live, Ross comes up with some other new ideas that I like. I particularly like his minimal way of relocating the asymmetry between the special sciences and physics: "the generalizations of the special sciences must not contradict those of physics, whereas

no symmetrical limitation holds in the opposite direction." And I find his ingenious way of using my idea of informational compression as the touchstone of ontology both novel and plausible. But I see a problem: I can't figure out how to fold into Ross's recipe what programmers call "*lossy* compression." Like Viger, Ross thinks I should abandon the abstracta-illata distinction. In "Real Patterns," I drew attention to the existence of patterns that were "imperfect," patterns that would have highly compressed descriptions if it weren't for the noise, the defects or blemishes. Idealized descriptions of those patterns describe *nothing,* strictly speaking, since they *over*simplify. But they impose a useful abstraction on messy reality: lossy compression of noisy data yields an abstractum. Or at least so it seems best to me to say. Those patterns, I claimed, are real, without being perfect, and the abstractum, a cognitive crutch of sorts, helps us see them. *We* see them, in spite of the noise, and so they are real for us—which threatens to lead to anthropocentrism. But so what? *Some* instrumentalist posits might indeed be of only local interest, if any. (Even I am not interested, really, in Dennett's lost sock center.) There couldn't be anything wrong with positing a few anthropocentric crutches. We *anthropoi* can make whatever crutches please us, creating *intentional* objects *ad lib.* Those that catch on, communally, create patterns that, although anthropocentric in origin, are in principle *visible* (even if baffling) to other beholders; they are patterns such that if "Martians" missed them, they would be missing out.

Two Black Boxes purports to describe just such a pattern. If you don't find an intentional stance explanation (there are several stylistic variants to choose from, depending on whether you are comfortable attributing mentalistic—not "just" semantic—properties to computers), you will be baffled by the near-perfect generalization that is visible to Martian and earthling alike: Pressing the α button causes the red light to go on and pressing the β button causes the green light to go on. Likewise (see the discussion of *Viger* above) lowering the center of gravity of the sailboat causes it to go faster. Some may want to say that strictly speaking it is not—could not be—the downward motion of the center of gravity (which is not to be token-identified with any particle) that actually causes anything, that we shouldn't confuse what we talk about in causal explanations with

what we take to be actual causes. I don't find this persuasive. To me, what count as actual causes are whatever we cite in explanations.

Ross has pointed out to me (in his editorial role for this volume) that even Dennett's lost sock center *could* play a causal role not so different from that of the sailboat's center of gravity. Suppose people need an arbitrary and neutral (in effect, pseudo-random) variable for some political purpose (e.g., deciding the order in which precincts shall get to vote) and hit upon using the wanderings of Dennett's lost sock center as their pseudo-random walk-fixer. Then if political factions ever figure out some political advantage to be gained by manipulating the order of voting, they can cause the order to change by causing Dennett's lost sock center to move (by causing me, in one way or another, to lose a few more socks in various locations). "'Twas the northerly motion of Dennett's lost sock center, you see, that caused the northern precinct polls to open first, and that caused the landslide." (Fill in the details as you like—I've given just the skeleton of Ross's imagined case.) Notice that what permits Dennett's lost sock center to make a causal difference, unlike the center of gravity of the sailboat, is its becoming an intentional object of communal note (like the gold in Fort Knox); causation does indeed run through the actual location of the lost sock center, but only if (and because) people are good at tracking its actual location so that their shared *beliefs* about its location exert the reliable effect at the next phase. This doesn't make its motion any less of a cause. Some causes produce their effects via beliefs; some don't.

Now is it conceivable that the pattern we *anthropoi* articulate in terms of beliefs could be described even more perspicuously in some currently unimagined Churchlandish terms? I *do* think that Churchland is right to *try* to find a better pattern than that found by the intentional stance (and we should try to find a simpler, better physics, too, while we're at it); I just don't think his hunch that he's going to succeed is remotely plausible. But if he succeeded, I guess I'd agree with him that it had turned out that all things considered, we should say (this is the diplomatic decision) that there really weren't beliefs and desires after all. The intentional stance could, I guess, come to be discarded as a myth, superannuated by the patterns created by its very articulation. In the same way it could turn

out, I suppose that there wasn't *ever* any gold in Fort Knox; it was all a hoax.

Rainforest Realism, if I understand it correctly, is my kind of realism indeed; it rules out only silly, unmotivated ontologies, but is otherwise remarkably pluralistic, tolerant of multiple "unreduced" levels of being, what Ryle was trying to get at with his different "logical tones of voice," so long as they can pay for themselves as patterns.

Realism and Consciousness: Thompson, Brook, Lloyd, Rosenthal, and Polger

Dan Lloyd's essay continues and expands on some the ontological themes of the previous section, but I have found it easier to lay the groundwork for a discussion of Lloyd's views by first looking at some of the other essays on consciousness. *David Thompson*'s comparison between Husserl's phenomenology and my heterophenomenology is illuminating in a number of ways, and perhaps I should add that this is no accident. My own philosophical training included a deeply influential dose of Husserl from Dagfinn Føllesdal when I was an undergraduate, and I have always had Husserl in mind—though more distantly remembered than assiduously reread—when working on my own version of phenomenology. Thompson exactly captures the main point of methodological disagreement between us: "Where Dennett's *Consciousness Explained* studies consciousness for its own sake while taking science for granted, Husserl investigates consciousness in order to establish a solid foundation for science." Note that *ideally* the two projects should converge on a common theory or set of answers—on anybody's view. I should expect that whatever assumptions about science I take on unexamined are innocent—they shouldn't require wholesale revision in the light of whatever theory of consciousness I arrive at, and they shouldn't blind me to important truths inaccessible from that vantage point—and Husserl should assume that his Cartesian starting point doesn't somehow start him off on the wrong foot and lead him to a distorted vision of science from which he can't recover. I, however, *suspect* that all Cartesian or "first-person perspective" starting points lure the theorist into inflated and distorted catalogues of the phenomena of

consciousness, creating whole genres of bogus *data* for the theorist to stumble over. Like Quine in *Word and Object,* I prefer starting with ordinary things and the science we can make of *them* and then extending the grasp of that vantage point upward (it *seems* to be inward) to such extraordinary things as conscious beings. Symmetry suggests a similar suspicion should run the other way, and so it does: the hue and cry over whether I'm "leaving something out" in my campaigns against qualia and the like, and Thompson's more useful observation that I risk leaving science "as a kind of skyhook without foundation."

Suspicions are not proofs, however. The way to vindicate Thompson's skyhook suspicion is to demonstrate that when, as Carr recommends, I "take phenomenology to its limits, namely to turn it back on its own position," I confound myself with contradictions. I have yet to see the case made in any particularity. Subject the scientists' own heterophenomenology as they conduct their experiments with color phi to the most painstaking analysis. Does it collapse? I don't think so. Hetero-heterophenomenology and its further offspring must avoid incoherence if my view is to be sustained, but I don't see any problems—yet. And, continuing the symmetry, I must say that if *any* Cartesian starting point manages to avoid the pitfalls of positing a *me*dium, it is Husserl's, because, as Thompson points out, Husserl anticipated—indeed helped to shape—my suspicion about the spurious ontology-fountains of Cartesianism. We are kindred spirits, and the *epoché* is our point of closest agreement, from either side of our first-person/third-person starting points. My "discreet charm of the anthropologist" is indeed the studied neutrality of Husserl's *epoché,* and I agree wholeheartedly with Husserl that, as Thompson says, "whatever intermediate entities there may be in the *process* of grasping an object, it is not these representatives that we are conscious *of,* but the object itself."

Thompson thinks, however, that I backslide into "representativism" in spite of this agreement. (Here he and *Lloyd* share a suspicion, but from different perspectives.) "Whether the experience is of something real or not, it is never about a brain event. . . . The notion of 'unwitting reference' is being misused here." I want to defend the notion of unwitting reference, for it is not quite the

simple idea—the simple mistaken idea—that Thompson supposes. It may be mistaken, but his objection misses the mark. Consider Peter Bieri's recent novel, *Perlmanns Schweigen* (written under the *nom de plume* Pascal Mercier). Since it is a fiction, a novel, it is not in any way about me, of course. I am not the intentional object constituted by any of its sentences. However, I have been told (not by Bieri, so this may not be true—I haven't checked), that there is a character in the novel who is modeled on me. The sentences about this fictional character do not *refer* to me, even if they may bear a noncoincidental, informationally rich, dependent relation to me. This is not reference, but it is the model for "unwitting reference" that I mean to exploit. Similarly, if some prankster in the jungle is the source of most or all of the beliefs about the feats of Feenoman, it is not strictly true that the Feenomanists' assertions expressing their creed *refer* to this man. Not quite. But almost. Similarly, it may turn out that when I claim to rotate a mental image in my mind's eye, there is something happening in my brain that is the source of most of the details I recount in my heterophenomenological narrative. Manifestly, I am not conscious of these brain processes, however imagistic they may be. Thompson says that Husserl would probably have no trouble countenancing such "brain-representations" and that's a good thing, since then Husserl and I can still be in agreement. These brain-images (if such there were) are part of the *hyletic phase,* part of the causal, material goings-on that make consciousness possible (on my amateur reading of Husserl). But just as the anthropologist may have a scientific interest in determining what is causally responsible for producing the curious contents expressed by Feenomanists, the neuroscientist may have a scientific interest in determining what is causally responsible for producing the curious contents of my "mind's eye imaginings." Notice that when there is, say, a table in front of me, *its* properties handily account for the content of my heterophenomenological declarations about my table-experiences, but when there is no table, when I am, say, hallucinating a table, the question arises of what, if anything, has properties that account for the content of these declarations. It might be something in the brain. It might not. There need not be anything, anywhere that has *just* those properties, but rather some conspiracy

of disparate causes, otherwise unrelated features of things and processes. And in any case, I agree with Thompson (and Husserl) that my intentional object in such a case of hallucination—what my experience is *about*—is *never* a brain process, however noncoincidentally the properties of some brain process may be linked to the details of my hallucination. (See *Lloyd* on this issue, and my comments above, as well.)

So when Thompson says "Brain events have no counterparts in the fiction and myth analogies," I think he overstates the case. Some (but not all) fictional characters have strikingly similar sources in the world; some (but not all) religious myths have real live sources in the world. These are not intentional objects, to be sure; but they are what I had in mind as the things to which "unwitting reference" was made. No doubt that is a poor term, in retrospect, but the point of calling it *unwitting* reference was to underscore the fact that it is not transparent to the experiencer (or the religionist, or maybe even the novelist) that there is such a source. We have no first-person authority, as Husserl and I both say, to the causes of our experience; that's why the epoché is such a good way of isolating the contents of experience from the contaminations of eager theorizing.

Andrew Brook, like many other philosophers, feels the seductive tug of seemings, and also of dichotomies that I have tried to dissolve, but unlike most, he doesn't just give in to his gut feelings and declare my view mistaken or crazy; he is circumspect in his allegiance, and his wary approach to my views is just what is needed. It exposes the problems. (Levine 1995, is another admirably forthright confrontation by someone whose bones tell him I must be wrong.)

(First, let me correct two misapprehensions that might be engendered by this mostly exemplary expression of my views. In the matter of S-O impasses, Brook oversimplifies my account. I grant that there are [large-scale] cases in which we *can* make a clear and principled distinction between Stalinesque and Orwellian. My point is that when you squeeze out the grounds for making these distinctions by reducing the time frame down to milliseconds, you squeeze out the *only* grounds for making these distinctions. He also uses the terms "filling in" and "finding out" in a way almost opposite to the way I would recommend. I would prefer to say that we don't bother

filling in our visual fields; we find out some of what's out there to be seen and more or less ignore the rest.)

Brook's analogy of the novel continues the themes of *Thompson's* comparison between Husserl's view and mine. A novel is a real physical object, made of paper and ink and such; there is *not* also a set of real but nonphysical objects "in between" the novel and the (partly factual, partly fictional) world it portrays—a world of phantasms created by the novelist's words. My position with regard to seemings is parallel: There are real physical bodies and real physical events in their brains that serve (in various roundabout ways) to project a fictional world, and there is *not* also a set of real seemings in between the brain events and the (fictional or real) world they depict. See also *Lloyd's* chapter on other ways of handling this temptation.

Brook discusses a contrast between eliminativism with regard to some category of putative things and "wanting a better model" of those things. This contrast is attractive to philosophers, for obvious programmatic reasons, but it is not anywhere near as crisp and clean as Brook *starts* presenting it (in due course he notes the problems—see note 5, which is excellent). How surprising does the "better model" have to be for us to declare that really, there are no such things after all? This has been a theme of mine since the example of fatigues in the Introduction to *Brainstorms* (xix–xx). The issue is political or expository, not factual. In the case of Brook's useful trio, the F level, the P level and the I level, the problem is that the F level is far from "pure"; it gets contaminated by the inadvertent self-theorizing we all do, introducing P level intrusions. Thus, consider what the F level is for framing a mental image: "I form an image in my mind [OK so far, but let's see how this image-talk gets unpacked]; it's a sort of picture [Oh? Where? How big? Is something looking at it?] or at any rate it's like vision somehow [Better; it's hard to fault *that* minimal claim; mental images do come in modalities—visual, auditory, tactile, and so forth] and it's more or less in my control, but there's only so much I can hold in my mind's eye at a time."

The F-level account of the concept of a subject illustrates this tug from the P-level, since to say a subject "operates by forming things like beliefs" is biased toward what we might call the hammer and

tongs view of belief formation. Compare it to "A subject is a locus of beliefs and desires." And "consciously focuses attention on tasks" suggests a captain, an agent deciding where to aim the searchlight next. Contrast it with "is constituted by the sequence of tasks in focal attention." Now there is nothing wrong with this F level stress on the voluntary agency of the subject—that *is* the idea of a subject—a central meaner, the Boss. But when we then look to the P-level theory, we must bracket that bias. (See *Thompson*'s chapter on whether Husserl goes this far with me; we may part company right here.) I say we must get rid of the subject (and the seemings) in the P-level theory. As Brook says, there seems to be a nasty dilemma. Only a theory of consciousness as the workings of a vacant automated factory—not a subject in sight—could be successful theory. If there is still a role for a subject, still tasks for a subject to perform, still a need for the subject as witness, then the theory is bankrupt from the start. And the same must be said of any leftover seemings. Seemings *to whom?* Seemings, as Colin McGinn tells us Frege insisted, have to be *somebody's* seemings, but if so, then we have to engineer seemings out of the picture as well. How? Not by denying the undeniable. Not by pretending that the phenomenology is other than it is, but by taking these two *persona non grata* and wedding them—you might call it dice-and-meld. First, the given must be broken up into lots of little microgivens, and the taking (by the subject) has to be broken up into lots of little microtakings of those microgivens. Then when you look at the model, you don't see any subject, and you don't see any single place where the given is taken; instead, all that work is parceled out into many little moments of content-fixation, and although these moments of content-fixation can feed others and so on indefinitely, we must not see them as preliminaries for some master taking yet to happen. (See below in the discussion of *Lloyd* on "is that all there is?" for a further set of moves that needs to be made before we can rest.)

"The difference between levels of pain and levels of pain tolerance seems to be perfectly real, certainly in many instances." Yes, in many instances. But let's not use that fact to support a metaphysical view that goes well beyond it. Compare it to the following issue. Is economic value real? Of course it is. Are things more expensive

now, or is it just inflation? (This parallel: Does it hurt him more now or is he just less tolerant of *that much* pain?) Yes, sometimes, when the background conditions can be held constant or tracked in ways we understand ("in 1960 dollars"), we can make perfectly good sense of this question, and answer it. Sometimes, however, the circumstances have changed so much that there is just no principled way of settling what would count as the correct answer. Does a live goat cost more or less today than it did in Julius Caesar's Rome? ("What is that in real money?" is a classic expression of naive realism about economic value. "What is that in real hurting?" is naive realism about pain.) Economic value doesn't have to be "intrinsic" or "absolute" to be real; neither does pain. But that means that a functionalistic, relativistic, nonintrinsic theory of pain does *not* leave out the ouch!

"In the case of the frog and most or all other cases like it, there is a quite determinate state of *something appearing*. It simply cannot be resolved what the thing appears to be like—not by us and, we can perfectly well allow, not by the organism either." I wonder if Brook would be as confident of this if we were to extend the range of his claim to *mechanical* frogs. Do things appear to them? I would be happy to work with such a concept of appearing, but I doubt it is what Brook thinks he is talking about here. In the end, I think Brook still refuses to join me in my rite of exorcism of the seemings that would *be* over and above the events that could happen even in a suitably sensitive and discriminating mechanical frog.[2]

Brook asks "What is it about some judgments, descriptions, in virtue of which they hurt, whereas others don't?" A good question. Let me try to answer it indirectly with a little story.

Once upon a time there was a guy named Dooley who complained about a judgment he kept making—it would only occur to him when he blinked, by the way. It was a judgment to the effect that somewhere at that very moment a dog was dying. That was certainly not a pleasant judgment, what with all the further reflections that invariably came in its wake, but it was not, you might insist, a *painful* mental event in itself. Well, wait till you've heard more about it. This judgment didn't just occur to him once

in a while. It occurred obsessively, for uninterrupted periods of several hours a day, whenever, during those periods, he blinked. While this judgment occupied his attention, all other thoughts were banished. It prevented him from concentrating at work and at play. It spoiled his mealtimes, and the very anticipation of the next bout of obtrusive judgment was itself enough to spoil his waking hours. And, of course, it released floods of neuromodulators and hormones, depressing his bodily functions, etc. Drinking helped. He found that when he was mildly drunk, the judgment that somewhere a dog was dying didn't seem to matter all that much. To hell with dogs—that was his usual drunken reaction. Not surprisingly, Dooley developed a drinking problem.

Question: Do these obsessive judgments *hurt* Dooley? I suppose not. But they certainly make him suffer. One would also suffer if one was given local (not general) anesthesia before being tortured, but nothing would *hurt* you during the torture. What is missing in both these cases, but present in the case of somebody who has just been kicked in the unanaesthetized shin, is a variety of *further* neuromodulator releases and neural firings that are apt to provoke/enable (1) identificatory judgments about a location of a particular feeling (these judgments can be wildly inaccurate guides to the location of any trauma, by the way), (2) intensified involuntary muscular spasms (though flinching in reaction to judgments might also be a feature of Dooley's predicament, and would be likely in the case of the anaesthetized torture victim unless blindfolded). I do not know whether local anesthesia diminishes the strength of torture as either a tongue-loosener or long-term behavior modifier. Needless to say, we are not likely to find out the answer to this empirical question, but we mustn't jump to conclusions. Anyway, that's what the difference is, I think, between cognitive events that hurt and those that don't.

Am I an eliminativist? I'm a *deflationist.* The idea is to chip the phenomenon of mind down to size, undoing the work of those inflationists who actively desire to impress upon themselves and everybody else just how supercalifragilisticexpialidocious consciousness is, so that they can maintain, with a straight face, their favorite doctrine:

now, or is it just inflation? (This parallel: Does it hurt him more now or is he just less tolerant of *that much* pain?) Yes, sometimes, when the background conditions can be held constant or tracked in ways we understand ("in 1960 dollars"), we can make perfectly good sense of this question, and answer it. Sometimes, however, the circumstances have changed so much that there is just no principled way of settling what would count as the correct answer. Does a live goat cost more or less today than it did in Julius Caesar's Rome? ("What is that in real money?" is a classic expression of naive realism about economic value. "What is that in real hurting?" is naive realism about pain.) Economic value doesn't have to be "intrinsic" or "absolute" to be real; neither does pain. But that means that a functionalistic, relativistic, nonintrinsic theory of pain does *not* leave out the ouch!

"In the case of the frog and most or all other cases like it, there is a quite determinate state of *something appearing*. It simply cannot be resolved what the thing appears to be like—not by us and, we can perfectly well allow, not by the organism either." I wonder if Brook would be as confident of this if we were to extend the range of his claim to *mechanical* frogs. Do things appear to them? I would be happy to work with such a concept of appearing, but I doubt it is what Brook thinks he is talking about here. In the end, I think Brook still refuses to join me in my rite of exorcism of the seemings that would *be* over and above the events that could happen even in a suitably sensitive and discriminating mechanical frog.[2]

Brook asks "What is it about some judgments, descriptions, in virtue of which they hurt, whereas others don't?" A good question. Let me try to answer it indirectly with a little story.

Once upon a time there was a guy named Dooley who complained about a judgment he kept making—it would only occur to him when he blinked, by the way. It was a judgment to the effect that somewhere at that very moment a dog was dying. That was certainly not a pleasant judgment, what with all the further reflections that invariably came in its wake, but it was not, you might insist, a *painful* mental event in itself. Well, wait till you've heard more about it. This judgment didn't just occur to him once

in a while. It occurred obsessively, for uninterrupted periods of several hours a day, whenever, during those periods, he blinked. While this judgment occupied his attention, all other thoughts were banished. It prevented him from concentrating at work and at play. It spoiled his mealtimes, and the very anticipation of the next bout of obtrusive judgment was itself enough to spoil his waking hours. And, of course, it released floods of neuromodulators and hormones, depressing his bodily functions, etc. Drinking helped. He found that when he was mildly drunk, the judgment that somewhere a dog was dying didn't seem to matter all that much. To hell with dogs—that was his usual drunken reaction. Not surprisingly, Dooley developed a drinking problem.

Question: Do these obsessive judgments *hurt* Dooley? I suppose not. But they certainly make him suffer. One would also suffer if one was given local (not general) anesthesia before being tortured, but nothing would *hurt* you during the torture. What is missing in both these cases, but present in the case of somebody who has just been kicked in the unanaesthetized shin, is a variety of *further* neuromodulator releases and neural firings that are apt to provoke/enable (1) identificatory judgments about a location of a particular feeling (these judgments can be wildly inaccurate guides to the location of any trauma, by the way), (2) intensified involuntary muscular spasms (though flinching in reaction to judgments might also be a feature of Dooley's predicament, and would be likely in the case of the anaesthetized torture victim unless blindfolded). I do not know whether local anesthesia diminishes the strength of torture as either a tongue-loosener or long-term behavior modifier. Needless to say, we are not likely to find out the answer to this empirical question, but we mustn't jump to conclusions. Anyway, that's what the difference is, I think, between cognitive events that hurt and those that don't.

Am I an eliminativist? I'm a *deflationist.* The idea is to chip the phenomenon of mind down to size, undoing the work of those inflationists who actively desire to impress upon themselves and everybody else just how supercalifragilisticexpialidocious consciousness is, so that they can maintain, with a straight face, their favorite doctrine:

the Mind is a Mystery Beyond all Understanding. They might be called hype-noetists. A lot of the work in *Consciousness Explained* (henceforth, CE) is just designed to undo the hype so that the job looks more do-able. And I note that in this task the book has rung up some significant victories. Not a few philosophers have been dumbfounded to learn that they were not conscious of as much as they thought, and that consciousness is not always an aid.

One of Brook's interesting questions is what the difference is between my account of the intentional stance and my account of consciousness. How do I explain the greater interest in real processes in the case of consciousness? Thus: A theory of consciousness, even a folk theory of consciousness, is already descending into the P-level, considering not just what it is rational for an agent to do given what it knows, etc., but taking the bolder tack of trying to figure out something about the actual processes involved, so that such tactics as *distracting attention* can come into play. As Brook notes, "focused consciousness makes a difference to performance." The tennis coach knows roughly what you, as an intentional system, believe and desire, but he appreciates that the wrong beliefs and desires are typically *influential* when you swing at the ball, so he devises stratagems for manipulating your consciousness. As Brook says, "performance is generally worse when conscious attention . . . is distracted, split between two tasks, etc." Generally, but not always. Sometimes consciousness gets in the way, and not just on the tennis court. I recently discussed this issue with Gary Burton, the jazz vibraphonist, who confirmed in his own case what I had often noted about my own jazz improvization: when soloing, "the main thing is for me to get out of the way"—by which he means that he *doesn't* focus consciousness on the notes he is playing, let alone what his hands are doing with the mallets. He may think instead, he says, about where he's going for supper, or whether the audience is hip enough for a certain sort of move, or whether he knows that man sitting in the corner.

Finally, Brook is right that I slight autobiographical (or episodic) memory in CE. I am setting out to repair that gap. The first installment is Westbury and Dennett, 2000.

Dan Lloyd wants to egg me on into anti-Cartesian territory I've skirted, and I am happy to follow his lead *part of the way*. But I do

have my limits. We all hate to give up our own oversimplifications, I guess. I think Lloyd may have given me some insight into how Jerry Fodor must feel when he reads some of my stuff: "Oh, no! Do I *have* to give up the Language of Thought? The alternative you sketch would make life ever so much more complicated!" Do I *have* to give up microtakings? I'm not sure that's such a good idea, and I am not persuaded by Lloyd's use of my own tool 3D, the slippery slope, against the boundary conditions for them. Powerful tools must be used with discretion, and my point about the temporal boundaries of events in consciousness was not that they don't have *some* boundaries (inevitably blurred if you look closely enough), but that they manifestly didn't have boundaries *within an order of magnitude* of the scale that some theories (or tacit background assumptions) would require. Yes, you *can* date the British Empire's becoming cognizant of the end of the War of 1812, but only to "the winter of 1814–15" or some such conveniently vague phrase. So I am not perturbed by his rhetorical questions about *just* when and where the microtakings have their onsets, since my theory doesn't demand microsecond timing or micrometer location for them.

Before getting down to the details, let me take this occasion to clarify and expand somewhat on his introduction into print of my private term "deepity." The etymological source was Joseph Weizenbaum, the computer scientist who created the Eliza program. Many years ago he told me about a remark his irrepressible thirteen-year-old daughter had made at the supper table the previous evening. He had delivered himself rather rotundly of a philosophical reflection, to which she had responded: "Wow. Dad said a deepity!" I delighted in the coinage, but went on to define it for my own purposes (in introductory philosophy classes) as an apparently profound observation that depends for this appearance on a subtle ill-formedness that lets it hover between a trivial truth and a whopping falsehood. What comes through to the unsuspecting is the illusory conjunction of truth and whoppingness. "Love is just a four-letter word" serves handily, since "'love' is just a four-letter word" is as true and trivial as "'salt' is just a four-letter word," while anybody who managed to think that love is a *word,* as opposed to an emotional state or an interpersonal relation, or whatever it is, would indeed

suffer from heroic misinformation and need most desperately to be set straight. I do commend the term to your use. There are deepities aplenty on the lips of our students, and Descartes is the progenitor (or godfather) of more than a few.

Now back to Lloyd's teetotaling campaign against representations. As a social drinker, I find a healthy place in my life for alcohol, and as a social thinker, I find a healthy place in my life for representations, and I think Lloyd is overlooking a benign home for mental representations after all—but I want to stress that I now think the only way to get there is by first endorsing Lloyd's abstemious resistance to *premature* representationalism. "Early" representations—posited, for instance, in *animal* cognition and in most human *perception*—is as dangerous an abuse as underaged drinking. Representations are, in effect, an adult preoccupation into which one is only gradually initiated. (My thinking on this has also been strongly influenced by John Haugeland's recent book, *Having Thought,* 1998, which builds social thinking on a representation-free base of the sort Lloyd is trying to develop. And see *Thompson* on Husserl on representativeness, a criticism with many points of similarity to Lloyd's.)

The great obstacle to such a view of the mind is, as Lloyd so vividly expresses, the hulking ghost of Descartes, and we have to put in place something like his *phenomenal realism* in order to keep the interiority of Cartesianism at bay as long as possible. (For a helpful alternative—but, I think, harmonious—perspective on this idea, see *Thompson*'s chapter for a discussion of Husserl's concept of *constitution.*) The idea of a "phenomenal complex" seems to me to do the trick, leveling the playing field, as Lloyd says. I have one misgiving to express, but I do so gingerly, since my ontological scruples are in disarray. I don't see why he insists on defining P-properties in such a way that their existence depends not on their being detected but on the brute presence somewhere "in the universe"—but even outside the lightcone, I gather—of P-detectors. This is an intensification—unwarranted, so far as I can see—of my not quite parallel distinction between *suspect* and *lovely* properties (Dennett 1991b):

> We do have a need, as Rosenthal shows, for properties of discriminative states that are in one sense independent of consciousness, and that can be

for that very reason informatively cited in explanations of particular contents of our consciousness. These properties are partially, but not entirely, independent of consciousness. We may call such properties *lovely* properties as contrasted with *suspect* properties. Someone could be lovely who had never yet, as it happened, been observed by any observer of the sort who would find her lovely, but she could not—as a matter of logic—be a suspect until someone actually suspected her of something. Particular instances of lovely qualities (such as the quality of loveliness) can be said to exist as Lockean dispositions prior to the moment (if any) where they exercise their power over an observer, producing the defining effect therein. Thus some unseen woman (self-raised on a desert island, I guess) could be genuinely lovely, having the dispositional power to affect normal observers of a certain class in a certain way, in spite of never having the opportunity to do so. But lovely qualities cannot be defined independently of the proclivities, susceptibilities, or dispositions of a class of observers. Actually, that is a bit too strong. Lovely qualities *would* not be defined—there would be no point in defining *them,* in contrast to all the other logically possible gerrymandered properties—independently of such a class of observers. So while it might be logically possible ("in retrospect" one might say) to gather color property instances together by something like brute force enumeration, the reasons for singling out such properties (for instance, in order to explain certain causal regularities in a set of curiously complicated objects) depend on the existence of the class of observers. . . .

On the other hand, suspect qualities (such as the property of being a suspect) are understood in such a way as to presuppose that any instance of the property has already had its defining effect on at least one observer. You may be eminently worthy of suspicion—you may even be obviously guilty—but you can't be a suspect until someone actually suspects you. The tradition that Rosenthal is denying would have it that "sensory qualities" are suspect properties—their *esse* is in every instance *percipi.* Just as an unsuspected suspect is no suspect at all, so an unfelt pain is supposedly no pain at all. But, for the reasons Rosenthal adduces, this is exactly as unreasonable as the claim that an unseen object cannot be colored. He claims, in effect, that sensory qualities should rather be considered lovely properties—like Lockean secondary qualities generally. Our intuition that the as-yet-unobserved emerald in the middle of the clump of ore is *already* green does not have to be denied, even though its being green is not a property it can be said to have "intrinsically." This is easier to accept for some secondary qualities than for others. That the sulphurous fumes spewed forth by primordial volcanoes were yellow seems somehow more objective than that they stank, but so long as what we mean by "yellow" is what *we* mean by "yellow," the claims are parallel. For suppose some primordial earthquake cast up a cliff face exposing the stripes of hundreds of chemically different

layers to the atmosphere. Were those stripes *visible?* We must ask to whom. Perhaps some of them would be visible to us and to others not. Perhaps some of the invisible stripes would be visible to pigeons (with their tetrachromat color vision), or to creatures who saw in the infrared or ultraviolet part of the electromagnetic spectrum. For the same reason one cannot meaningfully ask whether the difference between emeralds and rubies is a visible difference without specifying the vision system in question. (1991b, 40–42)

I think Lloyd should drop the insistence that P-detectors must *actually* exist for specific P-properties to exist. Why can't we just accept that there are a kazillion P-properties that exist but are of no interest to us at all, our having no reason to suspect that any P-detectors exist anywhere to bundle them disjunctively into potent generalization-precipitators? It seems to me that my lost sock center (Dennett 1991a) exists alongside the center of mass of the moon; ontology isn't always important; importance is important (*pace* J. L. Austin). That qualm aside, I think Lloyd's account of P-properties and P-detectors is wonderfully illuminating, and just what we need to start fending off the dread question from the crypto-Cartesians: *Is that all there is?* Yes, it's hard to get your head around this point—Lloyd points to a few lapses of my own—and there are certainly those who very vehemently don't want even to try to get their heads around it. Even those who, like *Brook,* are willing to give it the old college try find it an alien and uninviting prospect. So Lloyd's staged withdrawal program makes good sense. First, let me take Brook from qualia to judgments, but then quickly from judgments to microtakings (for the reasons I discuss above). Then let's have phenomenal complexes, but I don't see why microtakings can't be the left-hand side—even the "inboard" side—of phenomenal complexes, once we see what they *aren't*—namely, representations in any Cartesian sense. I am happy to rechristen microtakings as eddies[3] in Lloyd's "Heraclitan brainstorm of neural inflection," if this helps.

Anyway, Lloyd and I are agreed that these eddies are not enough *like* pictures, maps, symbols, sentences, and other uncontroversial representations to be called representations, in spite of their own sort of intentionality. But I think there are other phenomena that really are, in quite a strong sense, mental representations. If I do

long division in my head, there are representations of the numbers just as surely as there are if I do it on the blackboard. If I conjure up the visual appearance of my late neighbor Basil Turner, there is something briefly in the world that *is* like a photograph of him. (When I do this, normally, the intentional object of my mental activity is Basil Turner, not my "image" of Basil Turner—this is Husserl's point, as *Thompson* emphasizes.) These special cases are terrible models to inspire us when thinking of the basic elements of perception and cognition generally, but they are real enough, for all that. Our capacity, as Gregorian creatures (Dennett 1995), to *use* such representations, as tools for thinking (Dennett, forthcoming), is, I think, one of the features that distinguish most sharply our minds from simpler minds. Descartes was right about *something,* and not just about the relation between algebra and geometry!

David Rosenthal shows that his higher-order thought or HOT theory of consciousness "explains why we should feel a certain reluctance to classify particular cases [of timing in consciousness] as being Stalinesque or Orwellian," but I think it does not go far enough, and this obliges him to bite several bullets that are extremely dubious (to me). Here, then, is my latest installment in the continuing constructive back-and-forth that Rosenthal and I have been conducting for about a decade. (My earlier rounds are CE itself, and 1991b, 1993, 1994.)

The basic idea of my multiple drafts model or pandemonium model (more recently recast as the "fame in the brain" model in Dennett 1996b, 1998b) is that consciousness is more like fame than television; it is *not* a special "medium of representation" in the brain into which content-bearing events must be "transduced" in order to become conscious. It is rather a matter of content-bearing events in the brain achieving something a bit like fame in competition with other fame-seeking (or at any rate potentially fame-finding) events. But of course consciousness couldn't be *fame,* exactly, in the brain, since to be famous is to be a shared intentional object *in the consciousnesses* of many folk, and although the brain is usefully seen as composed of hordes of homunculi, imagining them to be *au courant* in just the way they would need to be to elevate some of their brethren to cerebral celebrity is going a bit too far—to say nothing of the

problem that it would install a patent infinite regress in my theory of consciousness. The looming infinite regress can be stopped the way such threats are often happily stopped, not by abandoning the basic idea but by softening it. As long as your homunculi are more stupid and ignorant than the intelligent agent they compose, the nesting of homunculi within homunculi can be finite, bottoming out, eventually, with agents so unimpressive that they can be replaced by machines. So consciousness is not so much *fame,* then, as *influence*—a species of relative "political" power in the opponent processes that eventuate in ongoing control of the body.

The main difference between Rosenthal's HOT theory and mine, then, is that in his theory, being conscious is not like being famous; it's like being *known by the King.* In some oligarchies, perhaps, the only way to achieve political power is to be known by the King, dispenser of all powers and privileges. Our brains are more democratic, indeed anarchic. In the brain there is no King, no Official Viewer of the State Television Program, no Cartesian Theater, but there are still plenty of *quite* sharp differences in political power exercised by contents over time. What a theory of consciousness needs to explain is how some relatively few contents become elevated to this political power, while most others evaporate into oblivion after doing their modest deeds in the ongoing projects of the brain. (Why is *this* the task of a theory of consciousness? Because that is what conscious events *do.* They hang around, monopolizing time "in the limelight"—but we need to explain *away* this seductive metaphor, and its kin, the searchlight of attention, by explaining the *functional* powers of attention-*grabbing* without presupposing a single attention-*giving* source.)

Fame is not like television, not a medium of representation at all. Consider the following tale. Jim has written a remarkable first novel that has been enthusiastically read by some of the *cognoscenti.* His picture is all set to go on the cover of *Time* magazine, and Oprah has lined him up for her television show. A national book tour is planned and Hollywood has already expressed interest in his book. That's all true on Tuesday. Wednesday morning San Francisco is destroyed in an earthquake, and the world's attention can hold nothing else for a month. Is Jim famous? He would have been, if it weren't

for that darn earthquake. Maybe next month, if things return to normal, he'll *become* famous for deeds done earlier. But fame eluded him this week, in spite of the fact that the *Time* cover story had been typeset and sent to the printer, to be yanked at the last moment, and in spite of the fact that his name was already in *TV Guide* as Oprah's guest, and in spite of the fact that stacks of his novels could be found in the windows of Borders and Barnes and Noble. All the *dispositional properties* normally sufficient for fame were in place, but their normal effects didn't get triggered, so no fame resulted. The same, I hold, is true of consciousness. It is a mistake (though a very tempting mistake) to think of consciousness as a medium of representation in the brain, such that getting represented in that medium, for however short a time, counts as being in consciousness, whether or not the representation therein leads to the normal *sequelae* of reaction and influence. Of course you can impose such a definition of consciousness by fiat, if you like, but if you do take this course, you will find the costs prohibitive.[4] You will not be able to give an account of the role of consciousness in memory, or in guiding behavior, and you won't be able to explain the difference between unconscious and conscious mental activities or states except by positing some mysterious extra property of the medium of representation you choose.

Ned Block recently[5] objected to this aspect of my view of consciousness, plausibly claiming that if I am a good Rylean—and I like to think I am—I should embrace just such a dispositional analysis of consciousness *so that* I can handle momentary but historically inert flashes of consciousness. Ryle would insist that, say, a cooling glass goblet on the glassblower's pipe might become brittle for a fraction of a second before being rewarmed in the kiln; there is nothing to stand in the way of momentary unactualized dispositions. Just so, but my claim is that consciousness is *not* like that! I venture the diagnosis, moreover, that Block's supposition that what he calls access consciousness *is* like that might be what seduces him into wanting more: creating the gratuitous category of "phenomenal" consciousness to cover for the disappointment one might feel in enjoying a merely dispositional kind of consciousness. Similarly, poor Jim might complain to his agent: "You call that *fame?* That might be a sort of *access* fame–Oprah and all that—but it sure didn't feel like fame to

me! I want *phenomenal* fame!" Jim was disposed to be, as one says, a phenom, but he didn't quite make it.

The strength of the fame-in-the-brain view comes out quite clearly, I think, if we raise certain tough questions for Rosenthal's alternative, the HOT theory of consciousness. Can you have two HOTs simultaneously? A hundred? If not, why not? Is there a bottleneck, a single higher-order-thought-thinker in there, with lower-order thoughts waiting in the antechamber for admission to the throne room? And if you can have many HOTs at once, what if one is much more influential than all the others? Are the contained thoughts all equally conscious, or is there one brand or strain or lineage of HOTS that is the real you? Rosenthal wants to distinguish parroting from nonparroting speech acts, and he claims that on my pandemonium model this is difficult whereas for his it poses no problem. Is that not because he has quietly posited a central meaner to do the meaning? I think that he underestimates the power of the pandemonium model to underwrite whatever distinction we need between parroting and nonparroting speech. The competition in the pandemonium model is far from "aleatory"; it is a competition of skill and relevance in which unity is approached.

A word that stands out like a sore thumb in (my reading of) Rosenthal's essay is the pronoun "we." For instance "As we mentally hear ourselves say what it is that we think, we find it confused or unclear, and so revise on the fly what we think. But by adjusting our words as we go, we get the right thought out." As I noted in my commentary on *Brook,* any theory of consciousness with such a subject still in the picture has a hostage; until it is discharged, we can't tell if there has been any real progress. Contrast the quotations from Rosenthal above with this from the novelist (and excellent phenomenologist) Nicholson Baker:

> Our opinions, gently nudged by circumstance, revise themselves under cover of inattention. We tell them, in a steady voice, No, I'm not interested in a change at present. But there is no stopping opinion. They don't care about whether we want to hold them or not; they do what they have to do. (Baker 1996, 4)

Rosenthal's subject is in charge, judging and editing and endorsing, the last functionary to sign off before public relations issues the press

release. Baker's subject is an ironic observer, swept along by influential opinions. Baker's "we," just as much as Rosenthal's, has to be dissolved into the uninhabited machinery in the brain's factory, but Rosenthal's is still apparently doing a lot of work. And it is work that needn't be done, in my opinion. I don't find myself agreeing at all with his claim that "our subjective impression is always that our speech acts exactly match in content the intentional state they express." Back in *Content and Consciousness* (1969), I enshrined just such an idea in my ill-considered "awareness line," but came to see that this was a mistake. (Nobody is more critical than a reformed sinner.)

These are not the only points at which I now see that I have misled Rosenthal, and he is not alone. It is my own mode of expression that is often the culprit. I have misled everybody, myself included, on "probes," for instance, and I don't know how to repair the damage, so I warn all to avoid the generous assumption that there is a coherent doctrine there to be teased out. (Note that I have avoided the topic of probes altogether since 1991, hoping to find a better way of making the points that still seem to me to reside there. Some of these points have found alternative expression in the interim; others may lurk in the murk, but don't hold your breath waiting for me to reveal them.) In the meantime, let me just say that I do *not* intend the probe distinction to be what Rosenthal calls a third-person constraint. After all, Robinson Crusoe on his desert island was conscious of myriad things, but nobody *else* needed to react to them. It is internal fame, not external fame that matters. At another point, Rosenthal says, "The key is the denial by first-person operationalism of any difference between mental states and their being conscious—between how things seem and how they seem to seem." These are two different distinctions, on my view. And I do say, as he quotes, that actually framing a report can create conscious content, but this isn't the only way that conscious content gets created. Finally, I was nonplussed by Rosenthal's suggestion that my view might be that speech acts "simply settle which of these states wins out in the competition for expression, rather than converting subpersonal events of content fixation into genuine intentional states." I do not see that any such conversion is possible or necessary, since I don't see the distinction. This is related to something else I don't see in Rosenthal's essay.

Throughout, he adopts without defense a supposition of a sort of content realism that I think needs defense (and can't get it). It would take a separate essay to do justice to my hunch (and his meticulous presentation of the view I am so dubious about), so all I can do at this point is wave a caution flag.

Thomas Polger rises bravely to my challenge to philosophers to defend zombies as a topic of adult discussion, and in so doing clarifies the issue for all of us, a contribution no matter who "wins." It is fair to say that I had simply not imagined some of the subtleties of zombie-doctrine that Polger exposes, which goes a long way to explaining the failure of my campaign to achieve victory—yet. His difficulty locating my position in his taxonomy of varieties of zombism is itself an interesting datum. I say that the concept of zombies is ridiculous, and he doesn't find this either clear or simple. But eventually he clarifies the task before him: "So to answer Dennett's challenge . . . one must show that a conception of consciousness that allows for the possibility of *functionally* identical zombies does not entail epiphenomenalism." This is epiphenomenalism in the Huxley sense, since I gather that Polger concedes that I am right about what he calls the "metaphysically strict" sense of that term: It *is* ridiculous.

I must begin by clearing up a confusion about Huxley epiphenomenalism engendered by Polger's discussion. Suppose for the sake of argument that the difference between a car with a carburetor and a car with a fuel injector is not *functional*—they accelerate the same, are equally fuel-efficient, etc. Polger says that I make a mistake in declaring in such circumstances that carburetors are causal-role (or Huxley) epiphenomena. As he says, carburetors are crucial functional mechanisms in the cars that have them, and fuel injectors are likewise. But this misses the point of Huxley epiphenomena. If the *difference* between having one of these and having the other is not functional, then it—the difference—is epiphenomenal in Huxley's sense. It is not a difference that makes a functional difference, but it is a difference that makes a causal difference. Just look under the hood: One reflects light entirely different from the light reflected by the other (and—to harken to Huxley's case—if you listen really carefully, one gives off a whistle that the other doesn't).

The zombie defender is free to think of conscious states, or processes, or events as mechanisms that have certain properties, among which that they are conscious. Those mechanisms have causal powers, but they are replaceable with functionally equivalent mechanisms that have some distinctive properties. Some mechanisms are conscious, others are not. Both have causal powers—they are not epiphenomenal.

But they *are* epiphenomenal in Huxley's sense—since ex hypothesi their differences in causal powers don't make a functional difference. Setting aside the term, though, I agree that the zombie defender is free to make this move, but he does so at the cost of removing consciousness from the sphere of human interest it currently (and rightfully, if understood correctly) occupies. Polger himself gets close to seeing the problem: "Why should we care whether consciousness is part of our system?" And he answers that there are "a variety of reasons, such as aesthetic or moral reasons. Conscious states may be replaceable with respect to our bodies carrying-on their cognitive duties, but we value having them." But *why* do we value them? The one feature of my challenge that Polger overlooks here is one that I thought he was promising to address: that the zombie-concept is *useful.* If the difference between being conscious and being a zombie is like the difference between having aluminum or plastic plumbing, why should it make a moral difference which one you have? Why should it be immoral to dismantle a conscious person without permission (and without anesthesia) but not immoral to dismantle a zombie without "permission"?

"Dennett calls zombies 'pernicious' because he thinks that the defender of zombies is committed to saying that these important considerations rest on whether or not a subject has some epiphenomenal quality. That would indeed be troubling." Right. I don't see that Polger has made any progress on dispelling that concern, whether we call it epiphenomenalism or not. Given his concept of consciousness, we plastic-brained folks have no moral standing (we're zombies), while our aluminum-brained cousins have the morally important causal property—even though there is no difference in capacity, talent, prowess, or susceptibility to suffering between us. (They suffer, we zombies just suffer$_z$—which doesn't count, for reasons not addressed.) Pernicious indeed.

But the philosophical challenge to the zombie defenders is more basic than this, and need not be put in terms of the moral embarrassments of the view. For suppose, with Polger, that we have two sorts of beings in front of us, of which we know the following (I guess God told us): The type A beings are conscious; they are equipped with "mechanisms that have certain properties, among which that they are conscious. Those mechanisms have causal powers, but they are replaceable with functionally equivalent mechanisms that have some distinctive properties. Some mechanisms are conscious, others are not." The type B beings are equipped with these latter, nonconscious but functionally identical mechanisms. There is no difficulty telling types A and B apart; half the beings have green brains that sworl to the left and make heavy use of acetylcholine, and half have red brains that sworl to the right and make heavy use of serotonin, etc. The trouble is that the labels, "A" and "B," have been removed from our samples and mixed up. Our only task is to examine the samples and determine which are type A and which are type B. What do we look for? What is it about *any* nonfunctional causal difference you care to describe that could *motivate* us to decide that it is the difference that goes with consciousness rather than with unconsciousness? Notice that I am not playing verificationism here. I am not demanding "criteria"; I am asking for the minimum: something, anything, that would give somebody the slightest good reason for preferring the hypothesis that causal property *k* goes with consciousness, not unconsciousness. It will not do, of course, to see which set of nonfunctional causal properties most closely matches *us,* because we are not at this point entitled to any assumptions about how widespread consciousness might be among normal *H. sapiens.*

Evolution and Ethics: Mooney

Brian Mooney proposes virtue ethics as a safe haven for my somewhat inchoate views on ethics. In particular, my distrust of the rule-bound "perfectionism" of both consequentialist and Kantian ethical theories finds an ally in Plato, he says, who condemns "pleonexic maximization" and endorses a "craft" vision of justice. This is a timely

contribution since I'd been wondering what I ought to make of virtue ethics, and now I have a much better idea of the prospects, thanks to Mooney's usefully maverick rendition of the issues.

Mooney also hopes to persuade me to go further than heretofore in accepting a pluralism of values: "A good act is better than an evil one but not clearly better than a rose, just as a loving act is better than a hateful one but not clearly better than a just one." We should not bother trying to calibrate all values on a single scale—even though in the real world we must often make decisions that require us to adjudicate conflicts between them. Craft, phronesis, will do the trick.

Virtue ethics, thus seen, challenges rule-fetishism much the same way embodied cognition and situated robotics challenge GOFAI or High Church Computationalism (Dennett 1986, reprinted in 1998d): You don't need representations of rules in your physical symbol system, presided over by a theorem-proving deducer; you just need know-how, embodied skill, talent. The motto of the virtue ethicists, in competition with Bentham's rhyming rules or Kant's Categorical Imperative, might be: "You're good! Fly by the seat of your pants." But then virtue ethics, like embodied cognition, for all its attractions is still just a negative view with a largely unfulfilled promise and a few tempting examples. Not as bad as "Listen to the Force," but not a *whole* lot better either, so far as I can see. What are virtues made of? How is the thinking of a wise person organized, structured, and implemented? These are real questions that consequentialists and Kantians have forthrightly if implausibly answered. Their answers have their problems, but at least they have something on offer. We still need from virtue ethicists a theory of what structures and processes make for virtue in the individual agent. Mooney recognizes this: "it is not just practical wisdom and judgment based on considerations of particulars, and their relations to universal considerations, but more importantly, the *cultivation of certain dispositions that undergird practical deliberation* [emphasis added]." What are these dispositions, and how are they to be cultivated? Even good aviators must be schooled before they can fly by the seat of their pants, and what they learn in flying school, like what children learn in the course of their moral education, is a set of habits of thought, ways

of framing issues, *practice* in imagining situations soundly, and these can be seen as tools or instruments of moral decision-making.

> Just as you cannot do very much carpentry with your bare hands, there is not much thinking you can do with your bare brain. (Bo Dahlbom and Lars-Erik Janlert, unpublished)

Rules are for fools. You don't really have to understand them (and you certainly don't have to understand their rationale); you just have to be able to follow them doggedly (Dennett 1983). The rule fetishist thinks *all* competence comes from assiduous application of rules—the fundamental assumption of GOFAI as well—and the virtue ethicist thinks at least some competence is more native, more biological, less intellectual. And surely this is right, but it is not clear to me what *ethical* consequences Mooney wants us to draw (some day) from such a realization. We are not born moral, and some of us are more brave, caring, morally imaginative, sensitive, stubborn, than others. These differences give some a head start toward becoming proper moral agents. What supplements do we provide for those less well endowed by nature?

Furthermore, parts of our native endowment, however traditional and emotionally attractive, may stand in need of reconstruction. As Nietzsche and others have wisely reminded us, that something is natural is only prima facie grounds for supposing we should endorse it. Mooney emphasizes the importance of "commitment" to a "beloved" (or to a cause, in the form of loyalty), and we need to recognize that the varieties of commitment that come naturally to us may not all be morally unproblematic in the end, however valuable they are to us as conversation-stoppers when we are confronted with tempting opportunities to defect. Robert Frank (1988) has hypothesized that the evolutionary rationale for our falling "madly" in love, establishing a *commitment* beyond the dictates of sound economic reason ("Honey, you're the healthiest, smartest, prettiest member of the opposite sex who has paid attention to me *so far,* and time is running out; let's get married"), is an abridgment of reason by passion that protects both parties. (See Pinker 1997, 417–419 for a brief exposition.) This hard-boiled justification makes biological sense, and there are certainly some contexts in which it makes moral and

political sense—resistance to betrayal and blind obedience do indeed have their times as virtues—but they are also the apparent source of much of the world's worst evil. There is no all-purpose path from either ancient tradition or the biologically natural to the ethically defensible, so virtue ethics is, at best, a part of what we need at the outset of a long project.

Notes

1. Descartes extends this pretext of fiction to his *Treatise on Man.* (See footnote 1, Cottingham, Stoothoff, and Murdoch 1985, 99.)

2. In his *Kant on the Mind,* Brook (1994) attempts to answer the question I raise in *Kinds of Minds* (65–66, 93–98): What do we need to add to (mere) sensitivity to get *sentience* (which I take to be a synonym or at least close cousin to what Brook calls *simple awareness*)? I claim that "we shouldn't assume there is a good answer" to this question, and it is instructive to see that Brook himself is acutely aware of the problems associated with his own attempt to provide an answer. He starts with my old *awareness2* notion from *Content and Consciousness:*

> *A* is aware2 that *p* at time *t* if and only if *p* is the content of an internal event in *A* at time *t* that is effective in directing current behavior.

He adds "memory, other dispositions, and nonpropositional objects" to his account, but in fact my definition already includes all these under its umbrella of "internal events effective in directing current behavior," and then he adds "sensible manifolds" but this is simply declaring the events also sentient without further ado. My "conservative hypothesis" in *Kinds of Minds* about the problem of sentience is that "there is no such *extra* phenomenon. 'Sentience' comes in every imaginable grade or intensity, from the simplest and most 'robotic' to the most exquisitely sensitive, hyper-reactive, 'human'" (97). In short, adding "sensible manifolds" is piling on the *figment* to no good end.

3. I gave a talk in Stockholm some years ago entitled "Eddies in the Stream of Consciousness," about the "temporal anomalies" of color phi, metacontrast, and Libet's effects. When it was published in Swedish translation, the title had become "Regissörer i medvetandets strömmar" (Dennett 1992). I asked the translator about the meaning of the word "regissörer" and he replied that the word "eddies" was unknown to him, and he had just assumed that it was a neologism of mine, a diminutive of "editor," so in his version, my multiple drafts model was implemented in the brain by hordes of *eddies* or *editorunculi.* (I wish I'd said that! You will, Oscar, you will—in Swedish!)

4. Rosenthal's note 29 claims that my view can't distinguish between not being conscious of something and "its coming to be conscious too briefly to affect memory time." He is right, but this isn't a bug, it's a feature.

5. In discussion of my presentation to the King's College London conference on consciousness, April, 1999.

References

Baker, N. (1996). *The Size of Thoughts.* New York: Random House.

Blackmore, S. (1999). *The Meme Machine.* Oxford: Oxford University Press.

Brook, A. (1994). *Kant and the Mind.* New York: Cambridge University Press.

Cottingham, J., Stoothoff, R., and Murdoch, D., trans. and eds. (1985). *The Philosophical Writings of Descartes* (in two volumes). Cambridge: Cambridge University Press.

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

Dalhbom, B. (ed.). (1993). *Dennett and His Critics.* Oxford: Blackwell.

Dahlbom, B. and Janlert, L-E. (unpublished). *Computer Future.*

Dennett, D. (1969). *Content and Consciousness*. London: Routledge, Kegal Paul.

Dennett, D. (1978). *Brainstorms.* Montgomery, Vermont: Bradford Books.

Dennett, D. (1983). Styles of mental representation. *Proceedings of the Aristotelian Society, New Series,* 83, 213–226, 1982/83. Reprinted in Dennett (1987).

Dennett, D. (1986). Cognitive wheels: The frame problem of AI. In Dennett (1998d).

Dennett, D. (1987). *The Intentional Stance.* Cambridge, Mass.: MIT Press. A Bradford Book.

Dennett, D. (1991a). Real patterns. *Journal of Philosophy* 88: 27–51. Reprinted in Dennett (1998d).

Dennett, D. (1991b). Lovely and suspect qualities (commentary on Rosenthal, The independence of consciousness and sensory quality). In *Consciousness* (SOFIA Conference, Buenos Aires), E. Villanueva (ed.), 37–43. Atascadero, Cal.: Ridgeview.

Dennett, D. (1991c). *Consciousness Explained.* Boston: Little, Brown and Company.

Dennett, D. (1992). Regissörer i medvetandets strömmar. *Framtider,* 11: 21–22, Institutet för Framtidsstudier, Stockholm. (Also published as: Eddies in the stream of consciousness, *Future Studies,* Stockholm, 1993.)

Dennett, D. (1993). The Message is: There is no *me*dium (reply to Jackson, Rosenthal, Shoemaker, and Tye). *Philosophy & Phenomenological Research* 53 (4): 889–931.

Dennett, D. (1994). Get real. Reply to 14 essays. *Philosophical Topics* 22 (1, 2): 505–568.

Dennett, D. (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life.* New York: Simon & Schuster.

Dennett, D. (1996). The scope of natural selection (reply to Orr 1996). *Boston Review* 21 (5).

Dennett, D. (1996b). Consciousness: More like fame than television (in German translation: *Bewusstsein hat mehr mit Ruhm als mit Fernsehen zu tun*). In *Die Technik auf dem Weg zur Seele,* C. Maar, E. Pöppel, and T. Christaller (eds.). Rowohlt, 1996.

Dennett, D. (1998a). *Kinds of Minds.* New York: Basic Books.

Dennett, D. (1998b). The Myth of Double Transduction. In *Toward a Science of Consciousness II, The Second Tucson Discussions and Debates,* S. Hameroff, A.W. Kaszniak, and A. C. Scott (eds.), 97–107. Cambridge, Mass.: MIT Press. A Bradford Book.

Dennett, D. (1998c). Preston on exaltation: Herons, apples and eggs. *Journal of Philosophy* 95 (11): 576–580.

Dennett, D. (1998d). *Brainchildren.* Cambridge, Mass.: MIT Press. A Bradford Book.

Dennett, D. (forthcoming). Making tools for thinking. In *Metarepresentation,* D. Sperber (ed.). Vancouver Series in Cognitive Science. Oxford: Oxford University Press.

Dennett, D. (unpublished a). Memes: Myths, Misunderstandings, and Misgivings. Chapel Hill Colloquium, October, 1998.

Dennett, D. (unpublished b). The Evolution of Culture. The first Charles Simonyi Lecture, delivered at Oxford University, February, 1999. Available on the web at http://www.edge.org.

Dretske, F. (1981). *Knowledge and the Flow of Information.* Cambridge, Mass.: MIT Press. A Bradford Book.

Frank, R. (1988). *Passions within Reason: The Strategic Role of the Emotions.* New York: Norton.

Gould, S. J. (1982). Change in developmental timing as a mechanism of macroevolution. In *Evolution and Development,* J. Bonner (ed.). Dahlem Konferenzen (Berlin, Heidelberg, New York: Springer-Verlag).

Haugeland, J. (1998). *Having Thought.* Cambridge, Mass.: Harvard University Press.

Holland, J. (1995). *Hidden Order: How Adaptation Builds Complexity.* Reading, Mass.: Addison-Wesley.

Holland, J. (1998). *Emergence: From Chaos to Order.* Reading, Mass.: Addison-Wesley.

Lenat, D. and Guha, R. (1990). *Building Large Knowledge-based Systems: Representation and Inference in the CYC Project.* Reading, Mass.: Addison-Wesley.

Levine, J. (1995). Out of the closet: A qualophile confronts qualophobia. *Philosophical Topics* 22, 23: 107–126.

Orr, H. (1996). Dennett's strange idea (an enlarged republication of Orr's review in *Evolution*). *Boston Review* 21 (4).

Pinker, S. (1997). *How the Mind Works.* New York: Norton.

Preston, B. (1998). Why is a wing like a spoon? A pluralist theory of function. *Journal of Philosophy* 95 (5): 215–254.

Quine, W. v. O. (1960). *Word and Object.* Cambridge, Mass.: MIT Press.

Ramberg, B. (1999) Dennett's Pragmatism. *Revue Internationale de Philosophie* 53 (207): 61–86.

Sperber, D. (1985) Apparently irrational beliefs. In *On Anthropological Knowledge.* Cambridge: Cambridge University Press.

Westbury, C. and Dennett, D. (2000). Mining the past for the future. In *Memory, Brain, and Belief,* D. Schacter and E. Scarry (eds.). Cambridge, Mass.: Harvard University Press.