Author's Response

Science, philosophy, and interpretation

Daniel C. Dennett
Center for Cognitive Studies, Tufts University, Medford, Mass. 02155

Kitcher & Kitcher see "high drama" in the tug of war between the realists and eliminativists trying to lure me into their camps, but a more important drama plays in the commentaries: the tug of war between cognitive science and, shall we say, "pure" philosophy. Here the clash of Two Cultures can indeed be seen, as **Cussins** suggests. Instead of joining either camp, I will try to play peacemaker, explaining, as I see it, the strengths and weaknesses of each camp to the other, alternating back and forth between them. Along the way, this will permit me to respond to most of the questions and objections raised; I will try to respond to the balance in the last section.

1. The siren song of science. *The Intentional Stance* (henceforth *Stance*) is, as **Newell** notes with mild regret, a book primarily for philosophers. It offers "increased conceptual clarity," he says, but virtually no "technical development." **Sloman** joins Newell in urging me to get with the program and stop frittering away my time on philosophers' puzzles. **Van Kleeck** observes, correctly, that intentional system theory, as I have drawn it, offers nothing to the cognitive psychologist beyond "a slightly streamlined version of folk psychology – more parsimonious than folk psychology, though no more predictive."

I endorse **Newell**'s aims, and hope to contribute to the project of constructing a more testable, empirical, scientific theory, but the purpose of *this* book is to establish the philosophical foundations for such a theory. I agree with just about everything **Sloman** says, and, as comments below will show, I have indeed aspired to a theory that makes the sorts of novel predictions that **Van Kleeck** calls for. Why then didn't I get on with it, and abandon philosophy for empirical theory construction? In part, because as a philosopher I felt obliged to respond to the prevailing perplexities of my discipline, but, more important, because I thought, and still think, that most cognitive scientists are underestimating the conceptual problems inherent in their projects. This is nothing to quarrel over; they run the risk of investing their careers in mainly hopeless research paradigms that get off slightly on the wrong foot (consider the debris of twentieth-century psychology and neuroscience for many examples), whereas I run the risk of wasting mine on quibbles that will turn out to be easily resolved in hindsight, once the millennial empirical theory is in place (and there are enough examples of that to give a philosopher pause).

Table 1. *Table of commentators*

1. Science, not philosophy: **Kirsh, Newell, Sloman, Van Kleeck, Dummett, Smith**
2. Realism and instrumentalism: **Churchland, Cussins, Dretske, Stich, Lycan, Newell, Rosenberg, Taylor**
3. Ethology: **Cheney & Seyfarth, Griffin, Premack, Dummett**
4. Evolution: **Amundson, Goldman, Kitcher & Kitcher, MacLennan, Roitblat, Rosenberg**
5. Other issues: **Amundson, Danto, Dummett, Harman, Lycan, Newell, Searle, Taylor,** EDITORIAL COMMENTARY

A point on which we should all agree, I think, is that "principled" isolation of disciplines is a recipe for fantasy. **Kirsh** provides a clear statement of this point in his valuable corrective of some overstatements in my position. Note that when Kirsh says we must solve the competence/architecture problem simultaneously (I agree entirely), he is making a *major* claim not only about why philosophers should do AI but also about why AI workers should take the brain seriously. And the best do. **Newell**'s William James lectures (1987) are a case in point. I have argued for a similar thesis in Dennett (1984a), in which I accuse some investigators in AI of making what might be called *the philosopher's mistake*: trying to do a completely "pure" competence theory.

It is instructive to compare **Newell** and **Dummett,** who offer similar criticisms, but with a big difference. Dummett thinks it is a mistake for me to search for the organismic contribution to the determination of belief – a task he characterizes as "armchair neurophysiology." This is a striking instance of a philosopher taking a very interesting conceptual problem – indeed, for me, *the* interesting conceptual problem – and kicking it downstairs (to the neurophysiologists!), leaving philosophers with precious little to do. Dummett goes on to suggest that if my account is "meant as a piece of philosophy" (as opposed to an empirical proposal) it is not satisfying. Dummett wants something "classifiable as a theory" just the way Newell does, but he wants it to be responsive to nothing but the intuitions of philosophers and the demands of logic. It would be a formalization of "philosophy's very own phenomena" as Newell calls them: propositional attitudes.

Newell asks: "Will this central role of the propositional attitudes go on forever? Is it a good thing?" and these are the right questions to ask. My answer is that we can't yet see whether an empirical theory of belief (or whatever is closest to belief in a mature cognitive theory) will have any use for classical or close-to-classical propositional attitudes, but that Newell himself, in his description of the knowledge level, seems to me to presuppose that something very like propositions (and hence propositional attitudes?) are the elements from which the knowledge level is composed: "Knowledge is the medium. Each primitive system consists of a body of knowledge, including knowledge of the system's goals. If knowledge is input at some time, then it becomes a permanent part of the body of knowledge of the system." In Newell 1981 he says: "There are no structural constraints to the knowledge in a body, either in capacity (i.e., the amount of knowledge) or in how the knowledge is held in the body. Indeed, there is no notion of how knowledge is held (*encoding* is a notion at the symbol level, not knowledge level)." What is the smallest unit of knowledge that can be "input at some time"? A proposition? When a "body" of knowledge includes knowledge "of" the system's goals, does this mean that the body consists of a collection of units (propositions) that are *about* (that compose knowledge of) those goals? Newell intends to keep this level as neutral as possible with regard to alternative schemes of implementation (the symbol level), while still constrain-

ing it as tightly as possible by considerations of the sort **Kirsh** mentions. In the meantime, he models these units of knowledge in what is essentially the predicate calculus, and it is far from clear (to me) whether this partitioning of knowledge is innocent temporizing or a subliminal commitment to suspect sententialist mechanisms after all. One thing is clear: Only a careful examination of the actual technical proposals, their implications and their alternatives, will answer this question.

For instance, **Newell** cites Levesque (1984) and Dietterich (1986) as examples of the sort of technical development he hopes to see from me. Dietterich's work has not yet reached me, but Levesque's essay is a good example of what can happen when you turn your back on the philosophical complications and get on with your science. First, the good news: He proves a distinctly nontrivial Representation Theorem showing the correctness of a first-order symbolic realization of an interestingly powerful knowledge base *with knowledge about its own knowledge* (so it can be queried about what it knows and doesn't know). He also offers a tempting formal model of *knowing who or what something is*, in terms of knowledge about equivalence classes of coreferring expressions (his treatment is consonant, I think, with Millikan's 1984 discussion of reference and "protoreference," ch. 12). But (the bad news) all this depends on several philosophically controversial choices of idealization or oversimplification. All increments to the knowledge base are modeled by a single operation, TELL, and all utilizations of that knowledge are funneled through a single operation, ASK. (Levesque is very clear about the difficulties of advancing anything "concrete" about the knowledge level without descending prematurely to the symbol level [p. 157ff], and explicitly draws attention to the oversimplifications he is adopting.) Thus the only way for perceptual experience to augment knowledge is by its TELLing the knowledge base various things (couched in first-order expressions) – what the frog's eye TELLs the frog's brain, with a vengeance. And ASKing the right questions at the right time – which some see as the very heart of the problem of intelligent use of knowledge – is left to the wisdom of external interlocutors and is not a part of the model. So, for various reasons given in *Stance* and my other writings, I think Levesque has (probably) made the wrong choices for oversimplifications, but I do not yet have any better ones to propose. Cognitive science is difficult.

I must make a similar disclaimer about **Smith**'s proposed recasting of my distinctions between styles of mental representation into standard distinctions in computer science. I endorse his translations, right down the line, but note that they involve a slight narrowing of focus. In a digital computer (running Prolog, say), my distinctions come to exactly what he says they do. In a brain, they come to something at least strongly analogous; but it is not possible to say just what they come to in advance of achieving the degree of comprehension of the levels with regard to the brain that we have with regard to the computer. Particularly important is the business about rules and representations.

As **Smith** notes, rules are explicit from the programmer's point of view (thank goodness, since otherwise programming would be well-nigh impossible), but since the rules are not in the database, the procedures they represent are not explicitly, implicitly, or potentially explicitly known by the program. Moreover, to say that these rules are explicit from the programmer's perspective is to say that there is actually a syntax of rules. What computer can be programmed to do can be thought of as selecting a (typically branching and looping) path through an astronomically large space of possible computer procedures. Getting the computer to do one's bidding is accomplished by putting together huge molecules of systematic, syntactic elements. It is not yet clear that there is any analogous *system* to the constraining of dispositional space in the brain. The brain is plastic, and it is clear that various conditions shape and constrain its "choice" of procedures; it is no clearer that there is a systematic, additive, or "molecular" way these conditions accumulate than there is in the case of selection pressures in species evolution. That is, if the shaping is done by the simultaneous and partly chaotic imposition of a multitude of selection pressures (instead of the piecemeal, serial, and strictly systematic contribution of lines of code), then there will be no procedural language expressing the procedures (except, uninformatively, at something like the microcode level). If, on the other hand, there is a "programmer's perspective" on the brain, then the brain's merely tacit knowhow can be *articulated*, but that is scarcely more than a fond hope at this point. "High Church Computationalism" (Dennett 1986) then can be seen as the "bean-bag genetics" of AI – a nice comparison, since one can be a card-carrying neo-Darwinian without being "guilty" of bean-bag genetics, and a card-carrying believer in strong AI without being "guilty" of High Church computationalism – witness the connectionists.

2. Realism, instrumentalism, eliminativism: The joy of sects? An onslaught of "impassioned (and to the nonphilosopher arcane) commentary" is correctly anticipated by **Cheney & Seyfarth,** who ask what joy there is, after all, in philosophical sects. What joy indeed? Whose colors shall I wear, which oath of allegiance shall I swear? Eliminative materialism (**Churchland, Stich**), instrumentalism or analytical behaviorism or verificationism or logical positivism (**Lycan**), or upper- or lower-case realism (**Dretske**)? Does it matter?

Yes, but not in the way it seems to matter to some philosophers. (I intend the ambiguity – it seems to some philosophers to matter in ways it really doesn't, and it seems to some observers to matter to some philosophers in ways that it really doesn't matter to them.) What matters is that everyone who thinks about the mind, in whatever discipline, is guided in thought by rather subliminal families of images, hunches, and analogies, and these can be as blinding as any prejudice, or as illuminating as hints from God. In short, they are powerful and they are typically unexamined – except by philosophers. We philosophers are the ones who are not embarrassed to expose these subliminal contents to explicit scrutiny, to see what powers they have, thanks to logical implication, to (mis-)lead the un-self-conscious thinker. When we call what-is-implied-by-a-family-of-ideas a *theory*, we court ridicule from empirical scientists, since our theories, our variously named *isms* arrayed genealogically in logical space, do not do the work of empirical theories. They are meant merely to be globally and locally coherent ways of

understanding all the relevant terms at once, before setting out to make substantive discoveries, predictions, explanations. Thus they might better be called theory-schemata, or theory-frameworks, or something like that.

The labels are useful, since once a body of more or less received analyses, extrapolations, objections, and rebuttals to a particular family of ideas has been developed, this body of work constitutes a genuine product of philosophical research that should be readily retrievable. The trouble is that no sooner does a "position" get definitively branded than someone thinks up a new variation on it that preserves what was right about it while obviating the standard objections. Thus *refutation by labeling* is usually an anachronistic and fruitless enterprise. To see this, just note the difficulties I have gotten myself into by letting my view be called instrumentalism, in spite of its differences from "classical" – that is to say, refuted – instrumentalism.

The debate over "choosing sides" that seems at first to drive some of the philosophers' commentaries is better seen, I think, as the exploration of analogies. I say that attributing belief is *like* attributing a center of gravity, and *like* interpreting an artifact as a realization of a particular Turing machine, but several philosophers have doubts about these analogies. What one always wants to know about an analogy is where the points of similarity and difference lie.

Dretske says he is a realist about centers of gravity, and here the labels are really getting in the way, for he nicely spells out the standard wisdom about centers of gravity, and I agree. That is, we agree about centers of gravity; I think beliefs are like that. Does he? If he does, then whatever he wants to call us, we should be called the same thing, but I think we still disagree. The key is what he calls stance-independent properties. (See also **Newell**'s first question, about stances versus systems.) There is always more to learn about things we are realists about: electricity or electrons, DNA molecules or chromosomes; but there really isn't anything more to learn about centers of gravity or Mendelian genes. It doesn't make sense to wonder if they will turn out to be composed of microparticles, for instance. They are *abstracta*, and what we can learn about them is just whatever follows from their combinatorial roles in our theories. That is not nothing, but it is only stance-dependent facts.

Compare facts of chess. One can wonder whether or not a particular board position is possible (= can be arrived at by a legal series of moves) in chess, but not (sanely) about whether rooks might be, as yet unbeknownst to us, chess-molecules made of bishops and pawns. Facts of chess are not about little pieces of wood or ivory. Facts about beliefs are not about transient states of various nervous systems. Of course any particular account of a particular chess game has implications about the spatiotemporal trajectory of various hunks of ivory (or for that matter, transient states of some nervous systems, if the game in question is being played by two experts without a board, just telling each other their moves). And any particular account of someone's beliefs will have some rather indirect implications about the physical changes occurring in various parts of their nervous systems, but the two are best kept distinct.

Dretske says, "It is true, of course, that descriptions of *what* a person believes and wants are not fully determined by . . . the observable data on which we base our intentional attributions." Does his use of "observable" here mean that he nevertheless thinks that these facts are fully determined by internal, (relatively) unobservable data? As I say in the final chapter, contrasting my views with Fodor's (1975; 1987), it first appeared as if Fodor was challenging Quine's (1960) indeterminacy thesis by claiming that what is inside the skull settled the questions (peripheral) behaviorism left unsettled. I have always claimed that going inside leaves things only marginally better: enough better to quell the intuitions of outrage that rise up when faced with Quine's claim, but not all the way better. Quine's fundamental point about indeterminacy survives. I can't tell from Dretske's commentary whether he would agree.

Dretske says that beliefs "*are* in the head (like a virus is in the body), but what makes them beliefs, what gives them the content they have, is (partly) outside the head – in those 'perfectly objective patterns' on which we all rely to figure out what someone wants and believes." Change "beliefs" to "rooks" and "in the head" to "on the board." Rooks are on the board (like a cup of coffee on the table) but what makes them rooks, what gives them their identity, is (partly) outside the board.

I think it is best to consider beliefs, like chess pieces, to be abstract objects, but this analogy escapes **Churchland.** He wonders in what respects beliefs are *abstracta*. Here are a few: There can be an infinity of beliefs in a finite brain; between any two beliefs there can be another; the question of exactly when one has acquired a particular belief can have no objective, determinate (interpretation-independent) answer (you weren't born believing that lawyers typically wear shoes, and you have believed it for some time – just when, exactly, did this belief get added to your belief store?); the list of beliefs required to "get" the Newfie-joke (*Stance*, p. 76) is not a list of salient, individualizable states or objects in the brain.

Another reason for considering beliefs as *abstracta* emerges in response to **Lycan**'s fifth objection (I will deal with the others later), which concerns the relation between beliefs "as conceived by folk psychology" and the "content-bearing states" I claim as the province of cognitive psychology. Why, he asks, would the two only polysemously intentional phenomena (he means: phenomena that are intentional in different senses) have anything to do with each other? Answer: because we use the folk notion, the normative notion, to *label* the discovered states. This is an old point. Recall McDermott's (1976) castigation of "wishful mnemonics" – overfacile labeling by AI practitioners. He was just right: Calling something in your model *the belief that* puts a very high standard on its behavior – the standard drawn, with some poetic license, from the folk notion in all its idealization, which is not just the sort of idealization one finds in the Ideal Gas Laws. (This is also my response to **Newell**'s first question about "stance vs. system" and to **Rosenberg**'s claim that while the center of gravity has a definition that assures it of an unmysterious ontological status and a precise calculational value hinging on the mechanical values that define it, the relation of beliefs, as *abstracta*, to *their* mechanical underpinnings "remains a mystery." I think we do know in principle how such *abstracta* are related to the appropriate *illata*. It is the relation of the knowledge level to what Newell would call the symbol

level(s), but this is a relation with many possible variant accounts. I don't endorse any particular account yet, as Newell correctly laments, but I have a program.)

Churchland thinks beliefs as *abstracta* will eventually be swept away (like the crystal spheres of Ptolemaic astronomy – another analogy) once the exciting demonstrations of connectionism yield a full-fledged alternative to folk psychology. I think this is vastly premature, however, in spite of my longtime connectionist leanings (see ch. 3 of Dennett 1969).

Stich agrees with me on this point: The import of connectionism in this argument is simply that it shows, as he says, that there are *lots* of straws afloat. As for his second point, he is right that I favor the third of his proffered alternatives: "the best theories will be connectionist [in some sense – at any rate, not sententialist], and that no plausible way can be found to reduce folk theories to connectionist theories." Stich finds my grounds for "small r" realism in the face of this prospect unappealing, but I am unmoved by his call to Eliminativism. The standard example of vestigial folk theory not taken seriously in everyday talk is "the sun rises" – which we all say and no one believes, taken literally. But I for one will go on not just talking of, but believing in, sunrises and sunsets, properly understood. My ontological stomach is that strong. I will unblushingly acknowledge that my ontology also includes such items as voices (Dennett 1969), haircuts, opportunities, home runs, and "Old McDonald Had a Farm," in spite of the fact that no known or anticipated scientific theory has a smooth reduction of any of these. Eliminativists may try to tell us about the brave new world to come in which all these outmoded posits will be forgotten, but to me, this is just a tactical decision about how to talk, and an undermotivated one. There are still mountains of valuable things to discover about the denizens of these "discredited ontologies." (These Californians do expose my Rylean roots, don't they?)

Are beliefs more like atoms (**Taylor**) than like Mendelian genes or centers of gravity? Are they more like viruses than songs or chess pieces? More like alchemical essences than logical states of Turing machines? The analogies do help; they alert us to oversimplifications and overlooked avenues in our thought, but after a while, diminishing returns set in. I recommend that all the analogies be duly noted and mulled over, and then I recommend that we move on to other topics. That means leaving uncompleted a traditional philosophical task: composing a counterexample-proof, *deny-this-on-pain-of-death* "theory" of belief (in 25,000 words or less). Until I learn of a large cash prize for such a production, I will leave it to others.

Before leaving the issue of what, "strictly speaking," we should say, let me point out that although I am in general agreement with **Cussins**'s analysis of the issues that divide Fodor and me, I wonder about the last step of it, which strikes me as needlessly provocative. "Strictly speaking, then, there is, for Dennett, no such thing as the science of psychology, as there is a unique science of physics." This is a possible, but remarkably strict, reading of what a science is. It would raise an interesting question about biology as a science, for instance. To ride a well-worn science-fiction hobbyhorse: Suppose we find life on another planet that is not carbon-based. When we do the science of that life, will it be biology? Or is biology just the *so-called* science of one (local) subset of realizations of the type *living thing*? Maybe there cannot be any such alternative realizations of living things, but I doubt biology should be seen as "committed" to that hypothesis. Note, by the way, that some highly abstract parts of biology might go over into the new science just about intact: population genetics, for instance, or even some relatively abstract "neural" net theory (but not neurophysiology!). I would say that neurophysiology is just as much a science as population genetics, even if the former turns out to have "only" a local application on our planet. I would similarly speak of various levels of psychology as sciences – without worrying about whether they apply to all intentional systems or only to the (organic, Earthly) varieties.

3. Animals, anecdotes, and experiments. Back to Earthly science. *Stance* claims to have something to offer to researchers in various aspects of animal intelligence, and three commentaries focus on this.

Cheney & Seyfarth suggest that my contribution is "methodological." I do not take umbrage at all; that is what I thought my contribution might be. I am happy to get the evidence of Kitui's deflating performance, since it offers a clear instance of using the leverage of the intentional stance to interpret the data: What Kitui would have to believe and do (if he were rational and if he truly did have the higher-order intention to deceive) is not what Kitui in fact does; so much the worse for the higher-order attribution.

Cheney & Seyfarth correctly point out that my hypothesis about the role of periods of privacy as a precondition for full-fledged language needs testing. I did not mean to suggest that I thought it was proven by my initial, informal observations. But I find their politicians' example unconvincing as grounds for doubt. It may be true that "the lack of privacy seldom prevents self-justifying explanations of one's own behavior from reaching Byzantine levels of intentionality, even if they are transparently incredible," but this is probably true only of a species that already has a highly developed language system. Once language gets established, then of course all sorts of intricate possibilities become opportunities, but the evolutionary demands are more exiguous at the outset. I agree that it would be premature to elevate the correlation of lack of privacy into a cause, and it is hard for me to imagine what experiments could shed even indirect light on this, but let's try to think of some.

The issue between **Griffin** and me continues to be just this: when to speak of consciousness. We agree on the value of studies of animal communication (see my contribution to Whiten and Byrne 1988, and my commentary on their *BBS* article [Whiten & Byrne: "Tactical Deception in Primates" *BBS* 11(2) 1988], for instance), but Griffin wants to elevate as the "best criterion of intention (despite the difficulties of applying it), whether the organism thinks consciously about what it is doing and about the choices it makes among the alternative actions available to it." The problem with this is perhaps best illustrated not in this *BBS* treatment, but in the *BBS* treatment of Ewert ["Neuroethology of Releasing Mechanisms" 10(3) 1987]. Griffin notes that "frogs do much

more than catch bugs," but what does he make of Ewert's analysis of prey-catching in toads? Does it show that toads are conscious or that they are more like thermostats after all? The virtue of my position on this is that one isn't compelled to abandon one's intentional stance just because one has found a way of treating the whole toad as a mechanism that does not in any compelling sense "think consciously about what it is doing and about the choices it makes among the available alternatives" (see Dennett 1987b, commentary on Ewert 1987).

Toad romantics (I do not mean to suggest that **Griffin** is one of these), may be dismayed by Ewert's (1987) successes, for they can be read as disproving (or at least strongly undercutting) the hypothesis that a toad is conscious, and thus, on Griffin's proposal, that it has "real" intentions. This is one way of parsing the phenomena, but then real intentions are going to prove much rarer than one might have thought, even in the explanation of quite devious and intelligent mammalian – and even human – behavior. I want to continue to consider toads real intentional systems – just like monkeys, robots, and people – because what they have in common is theoretically important, despite the huge differences (see **Dummett**'s discussion of human thinking, for instance). By my lights, the misguided anthropomorphism comes from imposing our intuitions about (our kind of) consciousness on other intentional systems.

Premack, similarly, claims it is "an instructive matter, and a reasonably straightforward one, to distinguish real intentional systems from not so real ones," and offers three criteria. His first is what I call second-order intentionality: A "real" intentional system is one that attributes states of mind to others (has beliefs about their beliefs, for instance). I have trouble interpreting his second criterion, which seems at first to be circular, but I think what he means is that it is a mistake to attribute beliefs and desires to beings that do not themselves attribute beliefs and desires: Hence there are no mere first-order intentional systems. If you can't have a belief about a belief, you can't have a belief at all. His third criterion expands on this: Crocodiles and their ilk have "informational states" and can benefit from some sorts of "learning" (and people have similar states), but people – and some primates – have in addition some other sorts of informational states (beliefs, in his sense, for instance). Later in his commentary, he points out the particular powers of these second-order states in their self-attributive uses: having beliefs about the reliability of one's own perceptual states, for instance, and withholding action on the basis of doubts (which are essentially second-order). So, if I have interpreted Premack correctly, his disagreement with me comes down to this: There is a theoretically important subset of informational states, the real intentional states, that are enjoyed only by those creatures that attribute such states to others. These intentional states are different enough from the other informational states that it is a mistake to lump them together (as one does when one adopts the intentional stance toward a crocodile or frog).

My reply is that the informational states of frog and crocodile still are intentional in the "classic" sense (they exhibit "aboutness"), and are robustly amenable to invocation in intentional-stance prediction and explanation, so I am not convinced that this is a mistake at all. I agree completely that there are important theoretical reasons for distinguishing the higher-order intentional systems, and within them, their higher-order intentional states, and more particularly still, their reflexive higher-order states (their attitudes toward their own mental states), and that the presence of such states is a central mark of higher intelligence. I have myself made these points, not only in Dennett 1976 and in chapter seven of *Stance*, but also in my discussions of the importance of self-monitoring (Dennett 1984b; 1986b; forthcoming a, b). But there are also other theoretically important distinctions – such as the distinction I mark between beliefs (of any order) and opinions (language-infected "thinkings"). If the issue is just which distinct class deserves the everyday label of "belief" ("on any proper analysis" as he puts it), we have a terminological difference only, which would be nice to believe, since I have no quarrel with **Premack**'s categories as theoretically important ones. As for his discussion of "subintentional systems," I think we also agree about everything but terminology; I don't see that his demonstrations of "subintentionality" are any different from my demonstrations of "zero order intentionality" (p. 246); in both cases, what had seemed at first to be explicable in terms of belief and desire turns out to have a deflated interpretation.

Premack claims that my version of things obscures the possibility of multiple levels of control. This would be most unfortunate, for Premack's example of the pointing behavior in chimps, and his hunch about what would be found in the analogous human case, are suggestive of fruitful lines of further research. I think the experiment with humans should be conducted to see whether any action potential in the arm or characteristic "pointing" motor strip activity occurs, as he predicts. [See also Libet: "Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action" *BBS* 8(4) 1986.] I would predict the same, but I would go further than Premack; I don't think "conditioned response" is the only mechanism that could produce such an inhibited near-act. I think that other mechanisms of subpersonal cognitive psychology (not beliefs in his reserved sense, but still intentionally characterizable processes) could be invoked. I view the intentional stance as an avenue for *suggesting* such hypotheses for testing, not obscuring them, so I hope Premack is in the minority in being misdirected on this score.

Finally, **Premack** complains about my interpretation of Premack and Woodruff (1978). I wish I hadn't used the word "force," which understandably bothers Premack. I wish I had written of their "scrupulous efforts to *ensure* that their chimps engaged in nonanecdotal, repeatable behavior." Then at least it would have been clear that I was not objecting to anything like coercion in the experimental design, but just to the repetition involved in the training, which invariably provides a behavioristic alternative hypothesis – though not always a very plausible one. The problem as it arises in Woodruff and Premack (1979) is described in *Stance*, pp. 253–54, and apparently Premack has no quarrel with my account of the shortcomings of that experiment (Sadie, the deceiver). But he is right that the videotape experiment with Sarah is different. It has, in fact, just the right features, as he says: "For an individual to consistently choose 'solutions' it must, given certain controls, first detect 'problems,' and a creditable interpretation of 'problem' lies exactly in the

states of mind that the individual attributes to the actor." With ingenuity and persistence Premack and Woodruff compiled a series of experiments that *virtually* ruled out the boring behavioristic explanations of Sarah's choices – but not quite.

As **Premack** says, Sarah had an enormous amount of training in match-to-sample tasks, and can be presumed to have become something of a virtuoso in finding the "correct" (rewarded) answer to such problems. When given a range of photographic alternatives as continuations to the videotape sequences, she almost invariably chose the right one first, but after the early round of experiments, a case could be made that the right choices were guided by mere association, not an appreciation of the actor's "problem." A subsequent round quite properly sought to disconfirm this by presenting her with alternatives that were all about equally associated with the prelude material, so that she was apparently faced with a much more exacting task of discrimination. For instance, in the problem of the shivering actor and the unplugged heater, the solution (a photo of the plugged-in plug) had to be discriminated from a set of other photos of the heater plug – not plugged in, or with the wire cut. Still she chose the correct continuation, but since the right answer in the more demanding test was the very same photo that she had already been rewarded for choosing in the earlier, less demanding test, there is a good chance she simply remembered what the right, rewarded, choice was.

The experiment would have been better had the more demanding tests been performed with entirely novel materials, and a further round of experiments more or less met this requirement. It would have been better still if the dependent behavior had been something of unforced environmental significance to Sarah, not performance on a heavily trained *type* of task. Nevertheless, Woodruff and Premack conducted just the sort of experiment I applaud, and I was wrong to lump it with other, less telling experiments. (Given my error, I suppose it is understandable how **Premack** could have been mistaken in thinking I recommend "field observations over experiments" – the whole point of chapter seven, and its sequel, is to recommend experiments that will tease out the equivocations of field observations.)

4. Evolution, function, content. Does evolutionary theory show that function is determinate? Can it show that content is determinate? I say no, provoking rebuttals from **Amundson, Goldman, Kitcher & Kitcher, MacLennan, Roitblat,** and **Rosenberg,** who among themselves reveal some telling disagreements, which I will attempt to sort out.

MacLennan claims that the "central" problem of intentionality is consciousness. He may be right, but for years I have been proceeding on the plan that the best way to make progress is to divide and conquer thus: first a theory of content (intentionality) and then (on its shoulders) a theory of consciousness. Having fashioned a new set of shoulders, I am now hard at work on the head again. Time will tell if this is the right order of business.

He also thinks a "causal theory of intentionality" will permit one to be *objective* about "natural functional meaning" after all, and I am ready to rebut this idea right now. He is right that an intentional system is a nonequilibrium thermodynamic system (see Dennett 1984, ch. 2). But he is wrong that this observation can settle the problems of interpretation. To see why, just imagine setting a contest among engineers to build artificial nonequilibrium systems. Each such product will resist entropy increases more or less well, under various conditions. Which of its moves are mistakes, which are good bets that happen not to pay off, which are short-term winners that are long-term losers? Knowing that these things are (whether or not they are *supposed to be*) nonequilibrium systems, with a causally explicable "tendency to avoid equilibrium," leaves all these questions of their design (their design rationales) unanswered, and without answers to these questions, the promised "functional meaning" will still elude the observer. The situation is no different with "natural" nonequilibrium systems.

Kitcher & Kitcher demonstrate this with their guppies, in spite of their intentions. The guppy example is supposed to exhibit a case in which careful biological research yields determinate (but complex) attributions of function. Why then do they refrain from concluding their tale by telling us exactly what the function (or functions) of those guppy spots is and is not? I do not doubt a word of what they say, but find it does not disagree with anything I hold dear. Let us look, as they recommend, at the details: Suppose that Endler (1983) has shown that the guppy spots' function is not simply to attract females; it is tempered by the simultaneous constraint that the spots not be too large, thereby attracting predators. Fine so far. And presumably even more patient exploration could reveal still further benefits (and costs) derivable by the guppies from their spots. In particular, what of the selective benefits derived by some guppies of attracting Endler's attention sufficiently to come under his protection as worthy experimental subjects? At what point would it become part of the function (presumably an entirely objective, determinate function) of guppy spots to appeal to Endler? I suspect that Endler would take steps to prevent any looming environmental innovation that would threaten the populations he studies; so, thanks to their spots, some subpopulations of guppies have already won a tremendous – if perhaps not yet manifest – advantage over their less interesting or less accessible cousins. The Kitchers will agree then, I gather, that in virtue of this detail, it is beyond interpretation that among *the functions* of some guppy spots is Endler-attraction?

More generally, I fail to see that I have been unappreciative of the power of evolutionary analysis when it is done well (as **Kitcher & Kitcher** charge), any more than I am unappreciative of literary analyses when they are done well. I just don't think that either sort, however well done, discovers some fully determinate truth about what things mean, or what their functions really, truly are. It is no argument against Quine's (1960) indeterminacy thesis that some translations are *obviously wrong,* and it is no argument in favor of functional determinacy that some functional attributions are *obviously right* (the eagle's wing is obviously for flying, and the eye for seeing).

Similarly, in response to **Kitcher & Kitcher**'s proposed fixing of rationality, I agree that under many (local) circumstances we can agree on what patterns of thought "will tend to produce true beliefs," but this leaves ra-

tionality quite unfixed. While we're on details, what do the Kitchers make of the Kahnemann and Tversky (1983) example I used as my parallel with the biologist? They claim that in their view "we will still be able to charge people with irrationality when their behavior fails to exemplify particular principles that are normal for them, their species, or their subgroup." But what Kahnemann and Tversky show, presumably, is that it is normal for people to make these irrational moves. So it is rational, because it is normal? I doubt that this is what the Kitchers mean.

The point of **Kitcher & Kitcher**'s defense of determinacy of function is presumably to help support their realism about beliefs: Is belief without determinate content "belief manqué"? I don't think so. And I don't think Quine's goal was the general "condemnation" of belief, but rather, as I say in the last chapter of "Stance," the characterization of it as an indispensable dramatic idiom.

Goldman is another, though, who resists this slide from objectivity, and who thinks that a proper understanding of biology and evolution will restore determinacy of both function and content. He focuses on my argument in chapter 8, which he rightly sees as the heart of my case, but he fails to see how the parts of my case fit together. First he offers two "scenarios" that purport to argue against me, but in fact argue against the view I reject (p. 319) as an analogue of the Intentional Fallacy. Then he notices this, and in "fairness" notes that I claim that content in artifacts depends on the artifact's users (which is close enough for our purposes here). He then claims that in his scenarios, "there are no users at all." But this begs the question against my position, for I claim that there *are* users: the genes themselves are the beneficiaries relative to which, in the end, all functional questions must be adjudicated, and there are genes (one gathers) in each of his scenarios. These genes, inhabiting their survival machines, rely on these artifacts to preserve them even into the next second. By the time such a "cosmic accident" has lived a second, it has had all the users it needs.

I do wish, by the way, that philosophers would stop conjuring up intuition pumps based on cosmic accidents. They don't show anything reliable, so far as I can see. Suppose, by cosmic accident, the words "Drink Pepsi" were to emerge from the alignment of a trillion comets in the sky (I'm trying to keep a straight face as I write this). Would it be an advertisement for Pepsi? The Pepsi people would be delighted, and no doubt would put the Coca Cola people out of business. So what? There is no tooth fairy. In the real world, complicated structures don't just happen by cosmic accident, and so, when one wonders about their function (or meaning) their genealogy is obviously highly relevant – but it *settles* nothing (to suppose otherwise is to commit the Intentional Fallacy).

Amundson misunderstands me on this point. He is most puzzled by my "deemphasis of the *causal mechanism* of natural selection." But, as just noted, I don't "deny the relevance" of causal theories; I just deny their sufficiency. Some antiadaptationists think they can get by with *just* the causal facts ("Instead of asking 'What is good?' we ask, 'What has happened?' The new question does everything we could expect the old question to do, and a lot more besides." Ghiselin (1983), quoted in "Stance" p. 278). This, I hold, is a big mistake. The facts about causal mechanisms permit one neither to do without functional interpretation (the behaviorists' and Ghiselin's mistake) nor to make determinate functional interpretations (the mistake made by **MacLennan, Kitcher & Kitcher,** and **Goldman**). Of course adaptationists are committed to there being actual, specific causal histories that support their interpretations of characters as adaptations, but their commitment is only to there being the "right sort" of selective history, and they don't have to cash that out nonteleologically (fortunately, since it is impossible).

Like **Kitcher & Kitcher, Amundson** attempts to say how the details of scientific practice in evolutionary biology refute my analysis, but I think he underestimates my position. I am quite willing to grant the biologists their capacity to identify and reidentify homologies. I am also willing to grant the psychologists the capacity to identify and reidentify tracts of nervous system, eyes, lips, phones (in the linguists' sense), stimuli (subject to the usual problems of stimulus definition and generalization that bedevil behaviorists) and behaviors (subject to the same problems). But what one behavior means from one occasion to the next is a problem, and so is what one homologous trait's function is – as my discussion above of Endler's guppies shows.

There is causation to be taken into account at all times by the psychologist, and by the biologist. Neither, however, has succeeded in telling a watertight causal story that licenses functional, or referential, interpretations. The "right sort" of causal relations sought in the various causal theories of reference is now beginning to reveal itself as transparently teleological – and none the worse for that. (Witness Fodor's 1987 Laocoon-like toilings to avoid this conclusion.) And the "right sort" of causal history to settle a functional attribution in evolutionary theory is equally elusive to evolutionary theorists. (Witness Gould and Vrba's 1982 toilings to say when adaptation differs from "exaptation.")

So I think **Amundson** is mistaken when he concludes that there is no mentalistic analog for the causal principle of natural selection. There are causal principles on all sides: "choices" and "decisions" and other little triggers and transductions and the like. Causation is everywhere to be seen in cognitive psychology, and it is generally "relevant" – but just what does it show? Does it give a "historical, causal explanation in support of the semantic ascent from physical to teleological ascriptions"? Certainly, but such ascent stops short of determinacy.

This was the point I argued for in the passage in Dennett (1969) that **Rosenberg** discusses, where the issue is the interpretation of Fido's internal states vis à vis a certain steak thrown out onto thin ice. I supposed the scientist examining Fido to have historical, causal knowledge of the sources of the events he wished to endow with content. "He has the following information: an afferent event of type A, previously associated with visual presentations of steaks, has continuations which trigger salivation and also activate a control system normally operating when Fido is about to approach or attack something," Rosenberg gets the gloss on my early account slightly askew. I don't say that we don't need to identify the description under which Fido views the steak. I say that there is no fact of the matter as to exactly which description Fido is laboring under, but we do in fact need to

settle on some such description in order to make sense of the behavior. It provides the rationale to give one rather than another reading – in an entirely familiar way. It is analogous to our use of raised-eyebrow quotes in explaining each other's activities: "The Senator dismissed the allegations of the 'Communist dupe' without further investigation." Perhaps the Senator never used those words, to himself or others, but we have categorized the sort of attitude he had, an attitude that explains his dismissal – or rather alludes to an explanation that we can all readily conjure up. This is a particularly efficient way of drawing attention (rightly or wrongly, of course) to a large family of attitudes, expectations, beliefs, discriminative abilities and inabilities, foibles, obsessions, and fears that can be seen to color the Senator's actions. The extensional characterization will "suffice" in one sense: It identifies the victim of the Senator's action, but it does not in itself "guide" the psychology in the way Rosenberg suggests. Or at least not sufficiently.

I agree that in principle one can go lightly on interpretation and speak rather abstemiously just of *registrations*, where substitution ad lib is the rule. But one pays a heavier price for this than **Rosenberg** acknowledges. What the tendentious labels do for you is just to allude (without specifying) to the sorts of contexts in which results can be expected. This comes out particularly clearly in the sort of weird puzzle case I discuss in Chapter 5, "Beyond Belief": Suppose Tom *registers* a heavily armed fugitive mass murderer shaking hands with him (p. 199). Whether anything interesting follows depends entirely on what Tom registers this agent *as*. The same goes for Fido, of course; if Fido registers the steak *as* a mere clump of debris, he will no doubt pass it by.

I think this is consonant with **Roitblat**'s remarks about abandoning certainty: If we view the registration mechanisms, the syntactic engines, as making unjustified leaps of "inference" to meanings, then we can say what those states mean, but not what they certainly mean. That is, "good enough" is the standard for meaning attribution, both for theorists peering in, and for the organisms themselves. I think then that what Roitblat calls abandoning certainty of meaning is what I call abandoning determinacy of content. When Roitblat says his hierarchical accounts "require no mysterious components, homunculi, or executives to operate" he is missing a point of mine. His higher-level structures *are* homunculi, and there is nothing more mysterious about homunculi than that. Finally, although he is right that my view is intermediate between the vitalists and the eliminativists, he seems to think I view the intentional stance as erring in attributing "too much" intelligence to species; I do not, however, view this as a useful error, but as the truth, properly understood.

5. Other issues, alphabetically by author. There is an abrupt dismissal by **Amundson** of my claim that the theory of natural selection is quite unable to discriminate natural from artificial selection by an examination of the products. I take this to be a little-noted but incontrovertible fact, and I would like to see Amundson show just how a biologist would make this determination.

Some twenty years ago **Danto** suggested that beliefs were sentences in the head, and today he thinks "sententialism is here for a long time," despite my criticisms (and those of others), which "pass harmlessly overhead." He misses the point of my example of the "thing about redheads"; he says that it is no part of Realism to restrict internal representations to beliefs or even sententially characterized internal states. Quite true; that is why I myself am a Realist *about representations*, as I often say. My point is that many Realists *about belief* are committed to the view that all content must be packaged in propositions (read: sentences) that can make the sort of content-contribution to the whole that a well-formed-formula can make if added as a new axiom to a set of axioms. This is the issue that worries me about **Newell** and Levesque (see above). What Danto's insouciance shows is that if you leave the empirical questions to the empirical researchers, as he advises, and the semantical problems to "semantical experts" (who are they, if not philosophers, by the way?) you are left with nothing much to worry about.

Dummett raises several objections not treated above, and also provides an excellent discussion of the ineradicable vagueness of belief. He regrets my ignoring (human type) thinking, and finds my view that human belief and animal belief are on a par "grossly implausible." But, as mentioned above in response to **Premack,** I reserve another term, *opinions*, for the states Dummett correctly points to as playing a critical role in human action and psychology. I think, by the way, that *Stance* would indeed have benefited from another chapter, a positive account of opinions and the way they differ from beliefs, a sequel to *Brainstorms*' "How to Change Your Mind," but alas no such chapter has been written. The interanimation of opinions and beliefs in the accounts we must give of the distinctively human pathologies of self-deception and akrasia, alluded to in Dummett's remarks, and also the shaping role of explicit or nearly-explicit *thinking* in the guidance of normal human behavior, is a topic for further research by philosophers, and by me in particular (in my account of the nature of conscious thought, a topic I am currently working on).

Dummett also accuses me of making "the common compatibilist mistake of thinking that the success of the intentional stance must always depend upon the possibility of a parallel explanation," by which I think Dummett means: a causal, mechanical explanation in terms of inner goings-on, a subpersonal explanation. This requires some unpacking. In "On Giving Libertarians What They Say they Want" (*Brainstorms*, ch. 15), I showed how an intentional stance explanation could go through successfully even when it was known that no *deterministic* mechanical explanation of the behavior could be had. This denies one "common compatibilist mistake," the mistake enshrined in Hobart's (1934) title "Free Will as Involving Determination and Inconceivable Without It." But this is not the common compatibilist mistake Dummett has in mind, I gather. What can it be? If it is the belief that the success of an intentional stance explanation is always in jeopardy of falsification by the confirmation of a strikingly *non*parallel mechanical explanation, I don't think this is a mistake at all. I argue in *Brainstorms*, ch. 12, "Mechanism and Responsibility," that if one gets confirmation of a much too simple mechanical explanation (e.g., our hero's brilliant "act" turns out to have been the mechanical effect of a brain seizure – cf. the demoting discoveries about vervet brilliance), this

really does disconfirm the fancy intentional level account. So I do not see what "common compatibilist mistake" is left for me to have made.

Finally, **Dummett,** the premier interpreter of Frege, says that Frege's view is quite antithetical to the positions I discuss in Chapter 5. I am in no position to disagree with him. (Since I claim Frege's concept of a Thought as the "backbone" of current orthodoxy about propositions, I guess I am fair game if current orthodoxy turns out to be not at all what Frege had in mind.) But I am puzzled by one of Dummett's claims: Frege's principle that thoughts are not mental contents "implies that . . . when I grasp a thought, *nothing in my mind* determines which thought it is that I grasp. If so, nothing in my mind, still less in my brain, is required to determine what object or objects I am thinking about." I had wondered if this could be an implication of Frege's attack on psychologism, but dismissed the idea. If this is in fact what Frege held, then I wonder why his brand of antipsychologism should be taken seriously after all. I must be missing Dummett's point.

Harman worries about whether there is a vicious circularity in my definition of the intentional stance, a worry shared by **Lycan.** I don't think there is, for the following reason. First, I do not, as Harman claims, "abandon" my earlier substantive claims about the nature of belief and desire when I grant, in Chapter 4, that "one way to discover" what someone else believes and desires is to imagine yourself in that person's shoes and ask yourself "what would I believe and desire under those circumstances?" **Stich**'s projectionist method of belief attribution (see pp. 343–34) is, for all I can see, coextensive with my rationalist method (it will yield the same verdicts) if used by someone who already has a good idea what beliefs and desires are. One can feign total ignorance on this score, and then claim to be baffled by the circularity of Harman's Theses I and IV taken together. But even for one who claims to be innocent of preconceptions about what beliefs and desires might be, it seems to me that I have said enough, in the course of the book, about the principles governing attribution to brush aside the suggestion that in the end my account fails to break out of the circle Harman and Lycan have described.

Lycan's first three objections have not yet been treated. He begins with a fine summary of my position, correctly identifying three major reasons for my instrumentalism, and then turns to objections, written in philosophers' shorthand, warning me (and other philosophers) by allusion of various presumed difficulties. Herewith, then, my equally efficient but probably equally hermetical replies.

1. I share **Lycan**'s low opinion of certain excesses of logical positivism, but I am not an across-the-board instrumentalist, and do not advocate it (see pp. 71–72). What label would Lycan place on automata theory? Is one an instrumentalist in holding that to be a Turing machine is to be robustly interpretable as a Turing machine?

2. **Lycan**'s brief allusion to "the zombie example," which purportedly refutes analytical behaviorism, must be baffling to the uninitiated: they can take it from him that such examples "work" or they can take it from me that they don't work, or they can check it out in the literature. Lycan (1987) is a good instance on one side, while Dennett (1982) is a good instance on the other.

The editor also seems to believe in zombies. According to the EDITORIAL COMMENTARY, the mind/body problem is the problem of whether there is a difference between (1) behaving *exactly as if* one were in pain and (2) in reality having pain; it is suggested that there is no counterpart in biology or physics. If the editor took his own emphasized words seriously, I doubt if he'd find my view that there is no difference all that puzzling. To say that something (our putative zombie) *behaved exactly as if* it had pain is to say that it responds to analgesics exactly as if it had pain, begs for mercy exactly as if it had pain, is distracted from pressing projects exactly as if it had pain, is unable to perform up to snuff on various tests exactly as if it had pain, and so forth. About such an entity (such a putative zombie) I am unembarrassed to declare: Yes, there is no important difference – indeed no discernible difference – between such an entity and an entity that genuinely has pain. The illusion of a difference here is traceable to the tricky word "behavior," which can be understood to mean something like "peripheral," "external," "readily observable" behavior. I am not now and have never been *that* kind of behaviorist. I am the kind of behaviorist that every biologist and physicist is. The biologist says that once the *behavior* of the pancreas or the chromosome or the immune system is completely accounted for, everything important is accounted for. The physicist says the same about the behavior of the electron.

Searle also makes this mistake about my behaviorism, when he insists that my third-person point of view compels me to "look only at external behavior," but remember that everything the neuroscientist can look at is also external behavior by this criterion. See also my reply to **Dretske** above.

3. I don't find **Lycan**'s third objection, about instrumentalism and usefulness, compelling. I don't hold that the antecedent clause merely says something about "the usefulness of our predictive practices" any more than I hold that the sentence "Had TMj gone into state k, it would have . . ." says something about the usefulness of our predictive practices when we call things Turing machines.

Newell asks several questions that I haven't answered directly or by implication above. He is right that there is a lot more of folk psychology than the parts he and I attempt to clarify, but I wonder if he isn't selling the knowledge level short in saying that it concentrates only on the "knowledge and goals" of agents. Surely he also aspires to provide sound models of human *wondering, anticipating, learning, forgetting, perceiving, intending,* and many other phenomena familiar to folk psychology (and anathema to old-line behaviorists). So do I.

Newell asks about the *mystery* philosophers still see surrounding "the semantic relation" and I think the reason he doesn't see it is that he has so far chosen to sidestep it in his own work; for instance, when SOAR solves a block-piling problem, there is no question of just which blocks (in the world) SOAR is thinking about – for SOAR is not thinking of any particular blocks in the world. Once that issue arises, however, the sketchiness at best and incoherence at worst of existing "procedural semantics" accounts begins to be worrying. Here, by the way, **Dummett** and **Newell** might meet on common ground.

I tried, in Chapter 9, to give a sympathetic, detailed

analysis of Searle's (1980) various claims about the Chinese room, and to say where I thought he had gone wrong. In his present commentary, **Searle**'s response to specific objections to his position seems to be simply to declare, with no argument or analysis whatever, that they are "irrelevant." I have carefully considered Searle's three specific claims of irrelevance, but found them all to be invalid. Others are invited to confirm this finding. The rest of the commentary seems to me to be an exercise in *refutation by caricature*. (A student of mine once came up with the marvelous malaprop: "by parody of reasoning" for "by parity of reasoning." The difference is admittedly hard for many to discern, but mature students of philosophy realize that any philosophical position worth taking seriously can be made to appear utterly ridiculous with a few deft strokes of deliberate oversimplification – it's child's play, in fact.)

Taylor holds that to animals (clams? ants?) some things really *matter*, and that my nonrealistic way with "mattering states" means that I never really talk about the mind at all. My view has the effect, he says, of "interpreting away the central phenomena of mind." I think he is close to being right about something important: A more properly biological cognitive science would indeed pay more attention to "affect" and "emotion," and would never let the question hide in the shadows for long as to why some things matter to various organisms (or artifactual systems). But I think Taylor's way of arguing for his claim is indefensible. "There are too many reasons to state them all here," he says, why my view is a "catastrophic mistake." A list would be appreciated. Taylor chooses to focus on pain, presumably the most accessible example of my catastrophic mistakes: "About pain, there is a 'fact of the matter,' beyond all interest relativity, and immune to any supposed 'indeterminacy of radical translation.'" Now I deny this, and, as Taylor notes, I have written at considerable length about it in *Brainstorms*, trying to show how this is an illusion. Briefly, is it as obvious as Taylor claims that there is (always) a fact of the matter about pain? Do clams feel pain? Our sort of pain? The sort that matters? Do we human beings feel pain under various sorts of anesthesia and analgesia? In *Brainstorms* I went to the details to show that what is "obvious" to the armchair philosopher is not at all obvious when you start looking at the strange and troubling empirical facts. Anyone who thinks Taylor's brute assertion is obviously right might read that chapter and then reconsider.

Does mattering matter to me? It certainly does. I have written a whole book on it (Dennett 1984) – a book with which **Taylor** has strenuously disagreed. He sometimes strikes me as one of those who think that the only way to protect humanity from the dead hand of science is to erect an absolutist metaphysical Maginot Line. It won't work. I'm already way into enemy territory with my own account of mattering. I do not "sidestep" the issues, nor do I apply the "science upsets common sense" doctrine in a "knee-jerk fashion." When I apply this doctrine (as for example in my discussion of pain), I try to deliver the goods.

ACKNOWLEDGMENT

I am grateful to Kathleen Akins, Nicholas Humphrey, and Bill de Vries for advice on this Response.

References

Amundson, R. (forthcoming) Doctor Dennett and Doctor Pangloss: Perfection and selection in biology and psychology. *Behavioral and Brain Sciences*

Au, T. K. (1986) A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. *Journal of Memory and Language* 25:104–22.

Austin, J. L. (1962) *Sense and sensibilia*. Oxford University Press.

Bennett, J. (1976) *Linguistic behavior*. Cambridge University Press.

Bever, T. G. (1984) The road from behaviorism to rationalism. In: *Animal cognition*, ed. H. L. Roitblat, T. G. Bever & H. S. Terrace.

Bradley, M. C. (1964) Critical notice of J. J. C. Smart's *Philosophy and scientific realism*. *Australasian Journal of Philosophy* 42:262–83.

Brentano, F. (1925) *Psychologie von empirischen Standpunkt*. Meiner.

Brooks, R. (forthcoming) Intelligence without representation. In: *Artificial Intelligence Journal*, special edition on Foundations of AI, ed. D. Kirsh. Elsevier.

Brown, P. & Jenkins, H. M. (1968) Auto-shaping of the pigeon's key-pack. *Journal of the Experimental Analysis of Behavior* 11:1–8.

Brown, R. & Fish, D. (1983) The psychological causality implicit in language. *Cognition* 14:237–73.

Burge, T. (1979) Individualism and the mental. *Midwest Studies in Philosophy* 4:73–121.

(1986) Individualism and psychology. *The Philosophical Review* 95(1):3–46.

Byrne, R. W. & Whiten, A. (forthcoming) *Machiavellian intelligence*. Oxford University Press.

Campbell, K. K. (1970) *Body and mind*. Doubleday Anchor.

Carey, S. (1985) *Conceptual change in childhood*. MIT Press.

Cartwright, N. (1983) *How the laws of physics lie*. Oxford University Press.

Cheney, D. L. & Seyfarth, R. M. (1985) Vervet monkey alarm calls: Manipulation through shared information? *Behaviour* 94:150–65.

(in press) Truth and deception in animal communication. In: *The minds of other animals*, ed. C. A. Ristau & P. Marler. Erlbaum.

Chisholm, R. (1956) Sentences about believing. *Aristotelian Society Proceedings* 56:125–48.

(1966) On some psychological concepts and the "logic" of intentionality. In: *Intentionality, minds, and perception*, ed. H. N. Castaneda. Wayne State University Press.

(1976) *Person and object*. Open Court Press.

Chomsky, N. (1980) *Rules and representation*. Columbia University Press.

Churchland, P. M. (1970) The logical character of action-explanations. *Philosophical Review* 79:214–36.

(1979) *Scientific realism and the plasticity of mind*. Cambridge University Press.

(1981) Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78:67–90.

(1984) *Matter and consciousness: A contemporary introduction to the philosophy of mind*. MIT Press.

(1986) *Neurophilosophy: Toward a unified theory of mind/brain*. MIT Press.

(in press) Reply to Corballis. *Biology and Philosophy*.

Churchland, P. S. & Churchland, P. M. (1981) Stalking the wild epistemic engine. *Nous* 17:5–18.

Clement, J. (1982) Students' preconceptions in introductory mechanics. *American Journal of Physics* 50:66–71.

Darwin, C. R. (1874) *The descent of man, and selection in relation to sex*. Lovell, Coryell.

Davidson, D. (1970) Mental events. In: *Experience and theory*, ed. L. Foster & J. Swanson. University of Massachusetts Press.

(1973) Radical interpretation. *Dialectica* 27:313–28. Reprinted in D. Davidson (1984) *Inquiries into truth and interpretation*. Oxford University Press.

(1975) Thought and talk. In: *Mind and language: Wolfson College lectures, 1974*. Clarendon Press.

Dawkins, R. (1976) *The selfish gene*. Oxford University Press.

Dennett, D. C. (1969) *Content and consciousness*. Routledge and Kegan Paul.

(1973) Mechanism and responsibility. In: *Essays on freedom of action*, ed. T. Honderich. Routledge & Kegan Paul. Reprinted in Dennett (1978).

(1975) Why the law of effect will not go away. *Journal of the Theory of Social Behavior* 5:169–87.

(1976) Conditions of personhood. In: *The identities of persons*, ed. A. Rorty. University of California Press.

(1978) *Brainstorms: Philosophical essays on mind and psychology*. Bradford Books.

(1980) Passing the buck to biology. *Behavioral and Brain Sciences* 3:19.

(1982) How to study human consciousness empirically: Or, nothing comes to mind. *Synthese* 53:159–80.

(1983a) Intentional systems in cognitive ethology: The "Panglossian paradigm" defended. *Behavioral and Brain Sciences* 6:343–90. (Reprinted as chapter 7 of *The Intentional Stance*.)
(1983b) Taking the intentional stance seriously. (Response to Griffin). *Behavioral and Brain Sciences* 6:384.
(1984a) *Elbow room: The varieties of free will worth wanting*. MIT Press/Bradford Books.
(1984b) Cognitive wheels: The frame problem of AI. In: *Minds, machines and evolution*, ed. C. Hookway. Cambridge University Press.
(1985) When does the intentional stance work? *Behavioral and Brain Sciences* 8:763–66.
(1986a) Is there an autonomous 'knowledge level'? In: *Meaning and cognitive structure: Issues in the computational theory of mind*, ed. Z. W. Pylyshyn & W. Demopoulos. Ablex. (Commentary on Newell, same volume.)
(1986b) The logical geography of computational approaches: A view from the east pole. In: *The representation of knowledge and belief*, ed. R. Harnish & M. Brand. University of Arizona Press.
(1987a). *The intentional stance*. MIT Press/Bradford Books.
(1987b) Eliminate the middletoad! *Behavioral and Brain Sciences* 10:372–74.
(1988) The intentional stance in theory and practice. In: *Machiavellian intelligence*, Ed. A. Whiten & R. W. Byrne. Oxford University Press.
(forthcoming a) A route to intelligence: Oversimplify and self-monitor. In: *Can intelligence be explained?*, ed. J. Khalfa. Oxford University Press.
(forthcoming b) Cognitive ethology: Hunting for bargains or a wild goose chase? In: *Explanation of goal-seeking behavior*, ed. A. Montefiore & D. Noble. Hutchinsons.
Dietterich, T. G. (1986) Learning at the knowledge level. *Machine Learning* 1:287–316.
Dretske, F. (1985) Machines and the mental. Western Division APA Presidential Address. In: *The proceedings and addresses of the APA*, vol. 59, no. 1.
(1986) Misrepresentation. In: *Belief*, ed. R. Bogdan. Oxford University Press.
(1988) *Explaining behavior: Reasons in a world of causes*. MIT Press/Bradford Books.
Endler, J. (1983) Natural and sexual selection on color patterns in poeciliid fishes. *Environmental Biology of Fishes* 9:173–90.
(1986) *Natural selection in the wild*. Princeton University Press.
Evans, G. (1980) Understanding demonstratives. In: *Meaning and understanding*, ed. H. Parret & J. Bouveresse. Walter de Gruyter.
Ewert, J.-P. (1987) Neuroethology of releasing mechanisms: Prey-catching in toads. *Behavioral and Brain Sciences* 10:337–405.
Fodor, J. (1975) *The language of thought*. Harvester Press; Crowell.
(1980) Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* 3:63–110.
(1981) *Representations*. MIT Press/Bradford Books.
(1983) *The modularity of mind*. MIT Press/Bradford Books.
(1986) Why paramecia don't have mental representations. *Midwest Studies in Philosophy* 10:3–23.
(1987) *Psychosemantics*. MIT Press/Bradford Books.
Fodor, J. A. & Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28.
Frege, G. *Grundlagen der Arithmetik* (Foundations of Arithmetic).
Gallistel, C. R. (1980) *The organization of action*. Erlbaum.
Gardiner, M. (1970) Mathematical games. *Scientific American* 223(4):120–23.
Ghiselin, M. T. (1983) Lloyd Morgan's canon in evolutionary context. *Behavioral and Brain Sciences* 6:362–63.
Gilovich, T. & Regan, D. T. (1986) The actor and the experiencer: Divergent patterns of causal attribution. *Social Cognition* 4:342–52.
Gould, S. J. (1980) *The panda's thumb*. W. W. Norton.
Gould, S. J. & Lewontin, R. (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society* B205:581–98.
Gould, S. J. & Vrba, E. S. (1982) Exaptation – a missing term in the science of form. *Paleobiology* 8:4–15.
Griffin, D. R. (1983) Thinking about animal thoughts. *Behavioral and Brain Sciences* 6:364.
(1984) *Animal thinking*. Harvard University Press.
Grossberg, S. (forthcoming) *Neural networks*. MIT Press.
Gyger, M., Karakashian, S. J. & Marler, P. (1986) Avian alarm calling: Is there an audience effect? *Animal Behaviour* 34:1570–72.
Haugeland, J. (1981) *Mind design*. MIT Press/Bradford Books.
Heidegger, M. (1962) *Being and time* (translated by J. Macquarrie & E. Robinson). Harper & Row.
Hobart, R. B. (1934) Free will as involving determination and inconceivable without it. *Mind* 43(169):1–27.
Husserl, E. (1931) *Ideas* (translated by W. R. Boyce Gibson). Allen & Unwin: Macmillan. (The German original, *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie*, was published in 1913.)
Jennings, H. S. (1906/1923) *Behavior of the lower organisms*. Columbia University Press.
Kahneman, D. & Tversky, A. (1983) Choices, values, and frames. *American Psychologist* 39:341–50.
Kaplan, D. (1968) Quantifying in. *Synthese* 19:178–214. Reprinted in D. Davidson & J. Hintikka, eds. (1969) *Words and objections*. Reidel.
(1980) Demonstratives. The John Locke Lectures, Oxford University.
Kirk, R. (1974) Zombies vs. materialists. *Aristotelian Society Supplementary Volume* 48:135–52.
Kitcher, P. (1985) *Vaulting ambition: Sociobiology and the quest for human nature*. MIT Press.
(1987) "Why Not The Best?" In: *The latest on the best: Essays on optimality and evolution*, ed. J. Dupre. MIT Press.
Konolige, K. (1986) *A deduction model of belief*. Pitman/Morgan Kaufman.
Kripke, S. (1982) *Wittgenstein on rules and private language*. Harvard University Press.
Lakatos, I. (1970) Falsification and the methodology of scientific research programs. In: *Criticism and the growth of knowledge*, ed. I. Lakatos & A. Musgrave. Cambridge University Press.
Levesque, H. J. (1984) Foundations of a functional approach to knowledge representation. *Artificial Intelligence* 23:155–212.
Lindauer, M. (1971) *Communication among social bees*. Harvard University Press.
Lloyd, J. W. (1987) *Foundations of logic programming* (2nd ed.). Springer-Verlag.
Lycan, W. G. (1987) *Consciousness*. MIT Press/Bradford Books.
(1988) *Judgment and justification*. Cambridge University Press.
(in press) Ideas of representation. In: *Mind, value and culture: Essays in honor of E. M. Adams*, ed. D. Weissbord. Rowman & Allanheld.
MacLennan, B. J. (1988) Logic for the new AI. In: *Aspects of artificial intelligence*, ed. J. H. Fetzer. Kluwer.
Marler, P., Dufty, A. & Pickert, R. (1986) Vocal communication in the domestic chicken: II. Is a sender sensitive to the presence and nature of a receiver? *Animal Behaviour* 34:188–93.
Marr, D. (1982) *Vision*. MIT Press.
McCarthy, J. (1979) Ascribing mental qualities to machines. In: *Philosophical perspectives in artificial intelligence*, ed. M. Ringle. Humanities Press.
McClelland, J. L., & Rumelhart, D. E., eds. (1986) *Parallel distributed processing: Explorations in the microstructure of cognition* (2 vols). MIT Press/Bradford Books.
McClelland, J. L., Rumelhart, D. E. & Hinton, G. E. (1986) The appeal of parallel distributed processing. In: *Parallel distributed processing: Explorations in the microstructure of cognition*, ed. D. E. Rumelhart & J. L. McClelland. MIT Press/Bradford Books.
McCloskey, M. (1983) Intuitive physics. *Scientific American* 248:122–30.
McDermott, D. (1976) Artificial intelligence meets natural stupidity. *SIGART Newsletter* No. 57. Reprinted in J. Haugeland, ed., (1981) *Mind design*. MIT Press/Bradford Books.
Millikan, R. G. (1984) *Language, thought and other biological categories*. MIT Press/Bradford Books.
Monod, J. (1971) *Chance and necessity*. Knopf. (Originally published in France as *Le Hasard et la Necessite*, Editions du Sueil, 1970.)
Morgan, C. L. (1894) *An introduction to comparative psychology*. Walter Scott.
Nagel, T. (1974) What is it like to be a bat? *Philosophical Review* 83:435–51.
(1986) *The view from nowhere*. Oxford University Press.
Newell, A. (1980) Physical symbol systems. *Cognitive Science* 4:135–83.
(1982) The knowledge level. *Artificial Intelligence* 18:81–132.
(1986) The knowledge level and the symbol level. In: *Meaning and cognitive structure: Issues in the computational theory of mind*, ed. Z. W. Pylyshyn & W. Demopoulos. Ablex.
(1987) Unified theories of cognition. The William James Lectures, Harvard University. (Available on videocassette from Harvard Psychology Department.)
Newell, A. & Simon, H. A. (1976) Computer science as empirical inquiry: Symbols and search. *Communications of the ACM* 19(3):113–26.
Nicolis, G. & Prigogine, I. (1977) *Self-organization in nonequilibrium systems: From dissipative structures to order through fluctuations*. Wiley.
Nilsson, N. J. (1980) *Principles of artificial intelligence*. Tioga.
Norman, D. A. (1981) Categorization of action slips. *Psychological Review* 88:1–15.
Parfit, D. (1984) *Reasons and persons*. Oxford University Press.
Perry, J. (1977) Frege on demonstratives. *Philosophical Review* 86:474–97.
(1979) The problem of the essential indexical. *Nous* 13:3–21.

Piaget, J. (1929) *The child's conception of the world.* Routledge & Kegan Paul.

Popper, K. & Eccles, J. (1977) *The self and its brain.* Springer-International.

Premack, D. (1988) Does the chimpanzee have a theory of mind? revisited. In: *Machiavellian intelligence,* ed. A. Whiten & R. W. Byrne. Oxford University Press.

Premack, D. & Woodruff, G. (1978) Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1:515–26.

Prigogine, I. & Stengers, I. (1984) *Order out of chaos: Man's new dialogue with nature.* Bantam Books.

Putnam, H. (1974) Comment on Wilfred Sellars. *Synthese* 27:445–55.

(1975) The meaning of "meaning." In: *Mind, language and reality* (Philosophical Papers, vol. 2), ed. H. Putnam. Cambridge University Press.

(1978) *Meaning and the moral sciences.* Routledge and Kegan Paul.

(1981) *Reason, truth and history.* Cambridge University Press.

(1983) Computational psychology and interpretation theory. In: *Realism and reason* (Philosophical Papers, vol. 3). Cambridge University Press.

(1986) Information and the mental. In: *Truth and interpretation: Perspectives on the philosophy of Donald Davidson,* ed. E. Lepore. Blackwell.

Quine, W. V. O. (1956) Quantifiers and propositional attitudes. *Journal of Philosophy* 50:177–86. Reprinted in Quine (1966) *The ways of paradox.* Random House.

(1960) *Word and object.* MIT Press.

Reichenbach, H. (1938) *Experience and prediction.* University of Chicago Press.

Roitblat, H. L. (1987) *Introduction to comparative cognition.* Freeman.

(in press) A cognitive action theory of learning. In: *Systems with learning and memory abilities,* ed. J. Delacour & J. C. S. Levy. Elsevier.

(in preparation) Monism, connectionism, and a hierarchical action theory of learning. (Invited chapter in *Systems that learn,* ed. J. Delacour.)

Roitblat, H. L., Bever, T. G. & Terrace, H. S., eds. (1984) *Animal cognition.* Erlbaum.

Roitblat, H. L. & Herman, L. M. (in press) Animal thinking. In: *Proceedings of the Third International Conference on Thinking,* ed. D. M. Topping, V. N. Kobayashi & D. C. Crowell. Erlbaum.

Romanes, G. J. (1883/1977) *Animal intelligence.* University Publications of America.

Rorty, R. (1982) Contemporary philosophy of mind. *Synthese* 53:323–48.

Rosenberg, A. (1986a) Intentional psychology and evolutionary biology. Part 1: The uneasy analogy. *Behaviorism* 14:15–27.

(1986b) Intentional psychology and evolutionary biology. Part 2: The crucial disanalogy. *Behaviorism* 14:125–38.

Rosenbloom, P. S., Laird, J. E. & Newell, A. (1987) Knowledge-level learning in Soar. In: *Proceedings of the American Association of Artificial Intelligence.* Morgan Kaufman.

Rumelhart, D. E., McClelland, J. L. & the PDP Research Group (1986) *Parallel distributed processing: Explorations in the microstructure of cognition.* MIT Press.

Ryle, G. (1949) *The concept of mind.* Hutchinson.

Searle, J. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3:417–57.

(1982) The myth of the computer: An exchange. *The New York Review of Books* June 24:56–57.

(forthcoming) Turing the Chinese Room. *Artificial Intelligence.*

Sechenov, I. M. (1863/1965) *Refleksy Golovnogo Nozga.* St. Petersberg. Translated: *Reflexes of the brain.* Originally published 1863. MIT Press.

Sellars, W. (1954) Some reflections on language games. *Philosophy of Science* 21:204–28. (Reprinted with revisions in Sellars 1963.)

(1956) Empiricism and the philosophy of mind. In: *The foundations of science and the concepts of psychology and psychoanalysis,* Minnesota Studies in the Philosophy of Science, vol. 1, ed. H. Feigl & M. Scriven. University of Minnesota Press. (Reprinted in Sellars 1963.)

(1963) *Science, perception and reality.* Routledge and Kegan Paul.

(1974) Meaning as functional classification: A perspective on the relation of syntax to semantics. *Synthese* 27:417–38.

Skinner, B. F. (1964) Behaviorism at fifty. In: *Behaviorism and phenomenology: Contrasting bases for modern psychology,* ed. T. W. Wann. University of Chicago Press.

(1971) *Beyond freedom and dignity.* Knopf.

Sloman, A. (1985) What enables a machine to understand? In: *Proceedings of the Ninth International Joint Conference on AI,* ed. A. Joshi. Los Angeles.

(1986) Reference without causal links. In: *Proceedings of the Seventh European Conference on AI,* ed. L. Steels, B. Du Boulay & D. Hogg. North-Holland.

(1987) Motives, mechanisms and emotions. *Cognition and Emotion* 1(3):217–33.

Smith, M. P. (1986) Sphexishness, epistemic bounds, and a priori psychology. In: *Proceedings of the Eighth Annual Conference of the Cognitive Science Society.* Elrbaum.

Stalnaker, R. (1984) *Inquiry.* MIT Press/Bradford Books.

Stich, S. (1978) Beliefs and sub-doxastic states. *Philosophy of Science* 45:499–518.

(1981) Dennett on intentional systems. *Philosophical Topics* 12:38–62.

(1983) *From folk psychology to cognitive science: The case against belief.* MIT Press/Bradford Books.

Van Kleeck, M. H., Hillger, L. A. & Brown, R. (in press) Pitting verbal schemas against information variables in attribution. *Social Cognition.*

de Waal, F. (1986) Deception in the natural communication of chimpanzees. In: *Deception: Perspectives on human and nonhuman deceit,* ed. R. W. Mitchell & N. S. Thompson. State University of New York Press.

Whiten, A. & Byrne, R. W. (1988) Tactical deception in primates. *Behavioral and Brain Sciences* 11(2):233–73.

Williams, D. R. & Williams, H. (1969) Automaintenance in the pigeon: Sustained pecking despite contingent nonreinforcement. *Journal of the Experimental Analysis of Behavior* 12:511–20.

Wimmer, H. & Perner, J. (1983) Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13:103–28.

Winograd, T. (1975) Frame representations and the declarative/procedural controversy. In: *Representation and understanding: Studies in cognitive science,* ed. D. G. Bobrow & A. M. Collins. Academic Press.

Woodruff, G. & Premack, D. (1979) Intentional communication in the chimpanzee: The development of deception. *Cognition* 7:333–62.