

AN INTERVIEW WITH DANIEL DENNETT

AUGUST 2002

Tufts University

EACH YEAR AN interview with a significant modern figure in philosophy is included in *The Dualist*. This year, Professor Daniel Dennett graciously agreed to answer questions provided by *The Dualist* and the Stanford Philosophy Department.

Daniel Dennett is Distinguished Arts and Sciences Professor, Professor of Philosophy, and Director of the Center for Cognitive Studies at Tufts University. His thinking has profoundly influenced the philosophy of mind and consciousness studies, as well as work done in the cognitive sciences, neuroscience, and artificial intelligence. He is author of over a hundred scholarly articles and several books, including *Consciousness Explained* (1991) and *Darwin's Dangerous Idea* (1995).

Dennett: Thanks for the excellent questions. I hope my answers will clarify things for your readers as much as the questions clarified things for me.

Krista Lawlor:

You agree with Richard Rorty that incorrigible self-reports are a "mark of the mental," and you go on to claim that we can imagine robots making authoritative self-reports. You disagree with Rorty about the source of this authority. For Rorty, it is a matter of linguistic convention. But in "The Case for Rorts" you claim that the justification for crediting the robot Cog with authority in its self-reports is "natural" and deeper than convention. I am attracted to the claim that authority in self-reports goes deeper than convention, but I am wondering about your positive view:

On your view, just what is the justification for ceding Cog authority?

On the one hand, you build the case that it could happen that Cog is reliable, and better placed than competing authorities about its states. This is plausible, but greater reliability could not be the source of Cog's incorrigibility. (Incorrigibility involves the unthinkability of one's reports being undermined by anyone else, not just the happy fact that one's reports cannot be undermined by anyone in the vicinity.) On the other hand, you also suggest that the justification for ceding Cog authority derives from the fact that self-interpretation constitutes the mental. (As you say, authority accrues to a self-report "...because the

*states and events those self-reports are about get their function, and hence their meaning, from the subject's own *take* on them.") This would perhaps guarantee Cog's self-reports are not open to question from others. But it is a very different story than Cog's simply growing ever more reliable about its own inner states. Which account do you favor? And how would the methodology of the Intentional Stance make room for self-interpretation as a fundamental means of meaning-constitution?*

As you say, for Rorty, incorrigibility (as the mark of the mental) is "a matter of linguistic convention" and for me the authority that makes this convention natural "runs deeper." Let me try to explain what I had in mind in "The Case for Rorts" and how it seems to me now that Rorty has responded. Rorty no longer favors speaking of a "convention" here, but I think he was right years ago that there is something *rather like* a convention that establishes incorrigibility: when we start treating somebody else (or something else) as "one of us," an interlocutor with a mind of its own, this status is something seldom *earned* or *demonstrated* but just normally *vouchsafed* without fanfare or reflection. For instance, it grows out of the gradual appreciation that a child has acquired the ability to speak her own mind, tell us what she wants and what she is thinking, and hence is somebody one might well ask—indeed *ought* to ask—before making any major decisions affecting her interests or welfare. With newly encountered adults, this is the default assumption, to be overruled only when we discover grounds of very serious impairment of faculties. So far, no real—as contrasted to fictional—robot has charmed its way into the charmed circle, but when the day comes, as it probably will in the not so distant future, the robot's declarations about its own "mental" states will have exactly the same status as yours and mine because the status will be an instance of the same "linguistic convention." This status depends on our drawing a dividing line between the mental and what might be called the near-mental. You are authoritative about whether you *feel feverish*, but not about whether you actually have a fever. We let a simple thermometer overrule you on the latter point. When you declare your preference for chocolate over vanilla, we respect the declaration and hand you a serving of the chocolate, even though we may well suspect that next time and thereafter you'll opt for the vanilla. The reason the phenomenon of self-deception can exist in human beings (but not, apparently, in any other species) is that our Rortian quasi-convention allows us to grant special evidential status to the verbal declarations that are then belied by the actions that speak louder than those words. (We don't just say your talker-thing is on the fritz; we *take you at your*

word—and then convict you of self-deception.) But the price we pay for this shifting of something rather like the burden of proof is that we put ourselves on a slippery slide to trivialization. You can't be *that* wrong, but you can be right about less and less and less. (See my "How Could I be Wrong? How Wrong Could I be?" a commentary on O'Regan and Nöe: "The Grand Illusion," *Journal of Consciousness Studies*, forthcoming.)

My reason for wanting to go beyond Rorty is that I don't think this set of sociolinguistic or anthropological—or just plain social—facts can just sit there, brute conventions without a rationale. ("It's just what we do; that's just the way we earthlings treat each other.") There has to be a rationale, free-floating of course, for our having fallen in with, and sustained over the millennia, this tradition. And when we consider what that rationale could be, we can see—I think—the answer to your challenge. As you say, incorrigibility involves more than mere contingent reliability (greater than the reliability of any other sources in the neighborhood); it involves the "unthinkability of one's reports being undermined by anyone else." How could we pass to this unthinkability? And would the passage itself be merely conventional? We have other cases of such a shift. Think of money, which also has a conventional element in it. As something (cowrie shells, gold coins, pieces of printed paper, cigarettes) passes from being a regularly relied upon medium of exchange—not as hard to carry around as goats, or as apt to spoil as cheeses—to a currency, a burden of proof shifts. If I owe you 5 shells, and hand you 5 shells, I have discharged my debt whether or not you happen to want to take delivery of those 5 shells at that time, whether or not you would find it more convenient to be paid in melons or eggs. If you don't accept this, you just don't understand what money is. "He didn't offer you a deal, a trade; he *paid* you, fair and square!" Similarly, if he tells you he's got a pain in his left knee, he isn't just manifesting a symptom of left knee injury; he's *telling* you where the pain is. Now for these understandings to "work" both the payer and the would-be payee have to understand what money is; and similarly, both the teller and the audience have to understand what telling someone where it hurts is. Not any old parrot, for instance, can accomplish this speech act—though I for one would take Irene Pepperberg's Alex's utterances on this score very seriously—as seriously as I'd take a small child's avowals, but not an adult's.

In any case, once the status of the noise-maker is raised to *interlocutor*, a subset of its noises are *thereby* raised to the status of heterophenomenological reports, and with them comes the incorrigibility that is built in. "Simply growing more reliable" is a part of the story, but it must be met on the other side by the appreciation of

this growing reliability, leading to the eventual “conventional” tipping of the scales that turns “reports” (like the highly reliable “reports” my laptop can issue about what is going on inside it) into reports, which may be no more reliable, but which get put into a different category.

Peter Godfrey-Smith:

What are your latest thoughts about the status of folk psychology? Do you think that the view in “Three kinds of intentional psychology” is still basically right? Have the arguments against the more behaviorist/instrumentalist aspects of your view moved you very much?

Also, do you feel vindicated by the problems and frustrations encountered by those, like Dretske and Fodor, who have tried to develop a more “realist” view of mental representations and their semantic properties? How do you see the state of the debate?

The trouble with the arguments against “the more behaviorist/instrumental aspects” of my view is that they have generally been addressed to caricatures (in which I have somewhat acquiesced, alas). And yes, of course I feel vindicated by the problems encountered by Fodor (and less so, Dretske) and others who have kept hunting for the touchstone of “real” content, “real” semantics. I told them so.

The basic question is: does content come first, or is it a resultant of some sort? And the answer is that of course it is a resultant. How could it not be? (It’s like life—the marvelous and sustained outcome of a huge conspiracy of Rube-Goldberg effects, not some primitive of the universe.) Anybody who thinks that content is more primary than life is simply a mystic, and anybody who thinks life is primitive is ignorant. When vitalism died, so did content essentialism. Or, it should have. “Theories” of content that suppose it to be some sort of primal feature of the world are breathtakingly strange, from my point of view, but I’ve come to appreciate that there are a lot of holdouts on this.

Fred Dretske:

Philosophers make mistakes. What is your biggest mistake? (By a mistake here I mean a view or position you advocated in print that you later came to believe was not only false, but seriously false.)

What a good question! I’ve thought long and hard about this, thanks to Fred. I don’t know which of my mistakes is the “biggest”—because there are so many dimensions on which to measure one’s mistakes. I’ll oversimplify and list two: one is long on content and the

other is long on method, but each is at least substantial on the other dimension.

1. In *Content and Consciousness* (p.87) I said:

Of all the common analogies used to describe the brain, the analogy of a community of correspondents (which is the inevitable suggestion whenever there is talk of codes and languages in the brain) is the most far-fetched and least useful. It has the disadvantage of merely postponing the central problem before us by positing unanalyzed man-analogues as systematic elements in that which we are trying to analyze, namely Man. The “little man in the brain,” Ryle’s “ghost in the machine,” is a notorious non-solution to the problems of mind, and although it is not entirely out of the question that the “brain writing” analogy will have some useful application, it does appear merely to replace the little man in the brain with a committee.

My skeptical campaign against “brain writing” (aka: the language of thought) continues in various forms to this day, but I soon came to realize that the idea of replacing the little man in the brain with a committee is not a bad idea; it’s one of the fundamental good ideas of cognitive science. Turning a dread homunculus into a committee of dischargeable homunculi is a fine way to dissipate the mystery in manageable clumps. Leaving “unanalyzed man-analogues” in your model *pro tem* is a well nigh indispensable crutch, and eventually, when we get to replace the many homunculissimi with machines, the Golden Age of cognitive neuroscience will be upon us. I corrected my mistake fairly early in my career, so some readers, not having encountered the early Dennett, might think that this shouldn’t count. Fair enough, but it was a big eye-opener for me. (What big mistakes have I made *lately*? Well, whatever they are, I haven’t seen the light yet. Maybe next year.)

2. My methodological goof was actually many related ones, and they continue to this day, and I find I don’t know exactly how to characterize them—and hence I don’t exactly know how to correct them either, not having a clear enough sense of *just* what is wrong in them. The general error is something like *misreading the audience*, and more particular versions of it are *underestimating one’s (perceived) radicalness* and, erring on the other side, *overstating one’s radicalness*. If doing philosophy is always a matter of walking one tightrope or another, then these are ways of falling off to the left and to the right. The easier one for me to describe is the former: underestimating one’s radicalness, or as I prefer to see it, underestimating the strength of dualist yearnings

and fantasies. To this day, in spite of more or less monthly revelations, I am *still* gob-smacked by the ubiquity, tenacity, virulence, and relentless attractiveness of dualist ideas to my fellow enquirers, who just keep falling for the same old song again and again and again! I am apparently much more immune to their seductive charms than my fellow philosophers and I marvel at how they can be so readily taken in. The unfortunate upshot of this is that I then simply cannot bring myself to compose and present the arguments that (it appears) it would actually take to turn people around on this topic. Heavens to Betsy! I don't want to insult my readers, so I tend to understate or underexplain my objections, and in general glide by the embarrassing anti-materialist bits, presuming people to have grown up enough to have discarded these ideas. Again and again I'm proven wrong. I am inclined to think that I suffer from a perspectival illusion, along the lines of overestimating the number of Volvos in the USA if you live in suburban New England. I spend more time with cognitive scientists than philosophers these days, and my kind of materialism tends to be taken for granted among those I work with. Sometimes I think that a certain proportion of the human population (it may be a recessive gene) is just constitutionally dualistic (in the sense of finding the appeal of dualism to be irresistible, a sort of topical imperviousness to arguments and general good sense.), and I surmise that these folks naturally self-select, ending up in the world's philosophy departments simply because they aren't comfortable anywhere else, not being "material girls in the material world" as Madonna has put it. And that is why, year after year, there is this dismal tide of dualists and mysterians, drawing life-saving succor from each other, and keeping this weirdly undermotivated set of doctrines alive. (As you can see, if my underestimation of dualism is a mistake, it's one I'm not close to being able to correct!)

Falling off the tightrope on the other side is going along with the gag and letting folks like Ned Block pin extremist labels on me ("instrumentalist," "fictionalist," "eliminativist"), and then spending the next twenty years or so trying to wean people away from the simple stereotype and even simpler refutation of my presumably wild and crazy view. A bit of advice from Uncle Dan: Always be skeptical of the refutation of *any* "ism" on the basis of a few general principles. And that includes the just-maligned-by-me dualism! Wholesale refutation almost never works. Dualism, or associationism, or behaviorism, or instrumentalism, or realism, or eliminativism, orÖcan mean so many different things, and so many intelligent people have defended so many subtle variants that you really should be suspicious of any argument that purports to do them all in in one fell swoop.

Take your Xisms as drawing attention to certain shared emphases,

even shared “propositions,” but expect the very same proposition (ahem: we still lack consensus on the identity conditions for propositions) to be at the center of quite astonishingly different positions. The refutation of one may well leave the other unscathed. In fact, it is almost certain to do so. Why? Because it was precisely in order to salvage the golden good from the leaden bad that the various variant versions were promulgated. It is so disheartening to confect a lovely non-standard brand of Xism, with all the strengths of Xism and none of the weaknesses (so far as one can see), and then see it dismissed as, of course, a brand of Xism, which succumbs, as everybody knows, to Standard Refutation #2. (I try to remind myself of this whenever somebody confronts me with what they describe as a new and defensible sort of dualism.)

Michael Strevens:

A theme in your writing is our use of different “stances” to understand and perhaps predict the behavior of very complex systems.† The “design stance” allows us to understand aspects of the configuration of biological systems and of artifacts; the “intentional stance” to understand and predict human behavior. Some related questions about stances:

1. Do you think that the elements of the design or the intentional stance have been improved at all, or are they in much the same shape that they were hundreds (or thousands, or millions) of years ago?

No doubt the intentional stance and the design stance have been improved over the years, in much the same way the physical stance has been improved since, say, Aristotle. After all, the design stance has been not just augmented but multiplied many times over by the addition of wave after wave of novel design structures. We now understand pumps and airfoils and switches and memory traces and steering and brakes and engines and deflectors and reflectors and amplifiers and Öso we have all these useful parts into which we can break designs. So it looks at first blush as if only the intentional stance has stagnated. But this isn't true, either. Yes, Homer had about as keen a sense of what motivates people, what justifies their beliefs, what is forgotten, and so forth, as anybody today, but he probably didn't have the *articulated* generalizations that we have. His own account of Ulysses and the sirens, for instance, is a paradigm of an intentional-stance *type* of predicament, like akrasia, or self-deception, or a Prisoner's Dilemma, but did people in his day categorize what they experienced into such types as readily as we do today? I find it plausible, in fact, that Homer

marked a major shift in human (self-)understanding, not quite as Julian Jaynes put it in *The Origins of Consciousness in the Breakdown of the Bicameral Mind*, but along the lines he opened up. Think of novice chess players learning what a *knight fork* is, and coming to be able to just *see* one in the offing. Having the term helps. I suspect that the human race has moved from novice to grandmaster in the intentional stance game, thanks to the Homers and Shakespeares, among others, who showed us the way.

Michael Strevens:

2. *How did we acquire expertise in the use of the stances in the first place? Did we figure it out ourselves? Did natural selection do it for us? Or is the story more complicated?*

We learned by doing, presumably. And since those who learned more swiftly had a significant fitness advantage over those who learned only with effort, a Baldwin Effect could drive the intentional stance into something of an instinct. (See my contribution “The Baldwin Effect: a Crane, not a Skyhook,” and Godfrey-Smith’s, among others, in *Evolution and Learning: The Baldwin Effect Reconsidered*, MIT Press, edited by Bruce Weber and David Depew, in press for November, 2002.) Autism shows that there is an innate contribution to normal intentional stance prowess. But children get lavish doses of practice, in pretend play, puppet shows, stories and—in the most recent biological instant—television. My untested informal observation (which can be tested, and should be) is that parents who are more outspoken articulators of intentional stance attributions probably have kids who are likewise gifted. As I have often pointed out, non-human animals such as apes may be, as Nicholas Humphrey has put it (1976, “The Social Function of Intellect,” in P.P. G. Bateson and R. A. Hinde, eds., *Growing Points in Ethology*, Cambridge: Cambridge Univ. Press, pp303-17), natural psychologists, but they never get to compare notes, and this absence of the benefits of a division of labor surely goes a long way to explaining their often shocking obtuseness when faced with intentional stance puzzles.

Michael Strevens:

3. *Are there other stances out there to be discovered? If so, do we have much chance of discovering them?*

Good question. When we let our imaginations wander, trying to

imagine other stances (the astrological stance, the Panglossian stance, Murphy's Law) we see that, try as we might, our imaginary new stances are apparently just special cases of the intentional stance—as if we tried to set up the case for the game-playing stance, the getting rich stance, the falling in love stance. Our failure of imagination here should not be taken to be authoritative. I don't claim to have a proof that there couldn't be other stances out there to discover; I'm just confessing my own inability to see what space is left to be occupied by another stance.

Manuel Vargas:

To what extent is the account in Elbow Room intended to be revisionist about ordinary ways of thinking about free will? Or, is the account supposed to provide an account of free will within the constraints of commonsense thinking about these issues? In the “varieties of free will worth wanting” parts of the book, the suggestion seems to be that a scientifically plausible (and otherwise adequate) account of free will may ultimately depart from something what we might call “the folk concept of free will.” At other points, (for example, in the analysis of “can”) the account seems to claim, and perhaps even rely upon, the idea that commonsense thinking is already adequate to the task of a theory of free will. Rather than being revisionist in any substantial way, the argument there seems to be that our ordinary, free will-relevant notion of “can” never had any commitments to it that stand in need of revision. So, how should we understand the book— is it supposed to be a vindication of commonsense or a proposal for the best way to revise it?

Elbow Room was not intended to be primarily either a defense of common sense or a defense of an improvement thereof. The point of the phrase “the varieties of free will worth wanting” is certainly to herald a challenge to traditional conceptions, a recognition of a possibility to be explored that some of the varieties of free will long the focus of philosophical attention are not really worth wanting in any case. But “common sense” is always, for me (and I think, for everybody, whether they acknowledge it or not), the default arbiter when no other grounds can be found. You ask an “either/or” question when it seems to me a “how much” question is what is called for: *Elbow Room*, and its forthcoming successor, *Freedom Evolves*, defend a substantial portion of common sense against what I consider to be philosophical distortion, while at the same time sprucing up common sense with a few overdue improvements. Since common sense consists in a hard-to-titrate mix of logical truth, super-reliable (and hence

obvious) empirical regularity, yesterday's science, and currently alluring ways of thinking, it is not the sort of thing one should want to vindicate without exception.

David Hills:

One of your earliest and most important teachers was the Oxford "ordinary language" philosopher Gilbert Ryle. This can seem surprising, since you've spent so much of your own career applying scientific results to philosophical problems, applying philosophical arguments to scientific problems, and making philosophically important science accessible to the general public — activities Ryle himself can sometimes seem dead set against. What do you think you've kept from Ryle's own thinking about the mind? And what do you think you've retained or reacted against in his conception of the nature of philosophical problems and philosophical method?

First of all, I think Ryle was gloriously right about the mistake one makes when (in my terms) you confuse personal level questions with sub-personal level questions, *the* category mistake. In general, I think his account of category mistakes led to some red herrings, but his diagnosis of the follies of *misplaced* mechanical (and reactionary "paramechanical") hypotheses was right on (Ryle, *The Concept of Mind*, 1949, London: Hutchinson). It is tempting to me to see most of the history of cognitive science in the last half century as a series of forays and skirmishes in which those who get this point (whether or not they articulate it, whether or not they have ever read, or even heard of, Ryle) are misunderstood by those who don't. (A nice opportunity to consider this tempting view can be found in Kevin O'Regan and Alva Nöe's target article, "A sensorimotor account of vision and visual consciousness," in *Behavioral and Brain Sciences*, 2001, and the commentaries.)

Ryle saw (brilliantly, intuitively, but not systematically) that in order to escape the mysteries, we had to turn the mind inside out, in a certain way, and forsake the quest for golden nuggets of *content* or *phenomenality* (or something like that) hidden in secret places amongst the machinery. How do the personal level events of folk psychology map onto the events of sub-personal level cognitive neuroscience? Ryle didn't try to address this question directly, because he knew he was master of only the personal level, but that still gave him plenty to say about what we *shouldn't* look for. So my introduction of the stances, for instance, can be seen as part of my effort to illuminate the indirectness Ryle identified but could give no positive account of. Ryle

wasn't *opposed* to scientific explanation of the brain and mind (except in some lapses); he just didn't have anything to contribute to it. He was very supportive and effective, by the way, in finding suitable scientific informants for me in Oxford, which started me down the path you identify.

David Hills:

When you wrote The Intentional Stance, at any rate, you viewed your conception of intentionality as close to that in Sellars. Sellars wanted to regard talk of mental states as meaning more or less what it would have meant if it had been introduced by means of a piece of deliberate theorizing, theorizing intended to help us explain regularities in human behavior, commonsensically described. He wanted to regard privileged access as consisting in the fact that we can train ourselves to announce the occurrence of mental state M in ourselves, without inference, under all and only the conditions in which this imagined psychological theory would enable third parties to infer its occurrence in us. He took it that when we work out this program in detail, the theory we're speaking of models thoughts on sentence tokens in a language and sensations on pictures. How much of this conception do you take yourself to have adopted in your own work? And how does your understanding of Sellars relate to your rejection of any Fodor-style language of thought? (Sellars often talks of thoughts as sentence-like, but he seldom talks of inferences as computation-like and rarely if ever mentions Turing. Is this a Good Thing or a Bad Thing?)

Ah me, this is a tough question. It has been years since I've grappled seriously with my very complicated reactions to Wilfrid Sellars' ideas. As I hope my answer to Krista Lawlor suggests, Rorty and I both—I think—imbibed from Sellars the idea that privileged access could have a naturalistic explanation that eschews inner observation in favor of something like training, and of course the proximal vehicles of that training (as contrasted with the more distal products, overt speech acts) would have to be sentence-like things. Not in mythical Mentalese, say I, but in one's native natural language, such as English. Opinions, I have called them—roughly, sentences deemed true. And who or what does the deeming? Not more sentences; it isn't "syntax all the way down." This is not to deny that there are (semi-)productive systems of (well, quasi-) representation in the nervous system aside from the systems that specifically deal with linguistically infected states such as opinions, but just that these more animal systems need not—and indeed should not—be modeled on sentences *or* on

pictures. (I have quite a bit more to say on this in Hugh Clapin, ed., *Philosophy of Mental Representation*, just published by OUP. And so do the other contributors to that remarkable book, the fruits of a workshop held in Maine in 1999 which by my lights exemplifies the *right* way to do philosophy. I will be very interested to see what other philosophers make of it.) Sellars' views on how to identify functional states (e.g., as playing roles rather like those played by public sentences) were tantalizing to me—they struck me as a defiantly personal level attempt to do the work of sub-personal theory, a way of flirting with very un-Rylean prospects—but in the end I was not convinced. This continues to be one of my most frustrating bits of unfinished business, to which I return periodically after tackling easier topics.

You ask if it is a Good Thing or a Bad Thing that Sellars seldom talks of inferences as computational and rarely if ever mentions Turing. Ah, but there is computational and then there is computational. GOFAI (Good Old Fashioned AI—Haugeland, 1985) misled a lot of people—not just philosophers, but physicists such as Roger Penrose—into thinking that any computational model of inference would have to be along the lines of resolution theorem-proving, a fundamentally axiomatic system of deduction more or less automatizing the predicate calculus or something similar. What I have called the Walking Encyclopedia (Dennett, 2001, “Things about Things,” in *The Foundations of Cognitive Science*, Joao Branquinho, ed. Clarendon Press, Oxford, 2001, pp. 133-143.). But Turing himself was wiser than that and had subtle ideas about how psychologically realistic processes of (something like) inference could be computationally realized in systems that weren't consistent and had no pretensions of consistency. Sellars was similarly wise to leave the implementation of the inferential (or quasi-inferential) roles of mental states wide open.

David Hills:

Your work on consciousness has included three important warnings: Don't assume there's any single special place in the brain or mind where all the information available to the person “comes together.” Don't assume that there's some single special moment at which a given external event becomes available to consciousness by “crossing a finish line” in the brain or mind. Don't confuse a succession of representations with a representation of a succession. How do these three warnings relate to one another in a division of critical labor? How do they relate to your more positive views about consciousness and cognition?

All three focus on different aspects of what I call the Hard Question (not the Hard Problem): “And then what happens?” (*Consciousness Explained*, p. 255) Theories that postulate a time of consciousness (S is conscious of A at time *t*) but neglect the task of characterizing the work that then gets done, the effects that would not have occurred without that purported consciousness at time *t*, tend to encourage fallacious thinking about the contents of consciousness, such as the thought that *since Tom was conscious of A before he was conscious of B, Tom experienced A as preceding B*, which does not follow at all. Our lazy imaginations fall back into thinking of the “stream of consciousness” as a single, well-ordered sequence (in objective physical time) of “finished” contents (all interpreted, with their meanings fixed), which we *then* somehow appreciate or understand, but the appreciating is itself part of the interpreting of course, part of the work that must get done by processes distributed in space and time in the brain. To put it with nearly comical crudity, there is no reason why you can’t have the subjective experience of B-followed-by-A thanks to a process that first ‘appreciates A’ and then ‘appreciates B’ and finally ‘appreciates’ A to be the successor, not predecessor, of B. No further pantomime, no rendering, no showtime, in which first B appears on stage, and then A appears on stage, is required for this to be the eventual subjective take-home message. As I note in “Are we Explaining Consciousness Yet?” (*Cognition*, **79**, 2001, pp. 228-9):

A model that . . . undertakes from the outset to address the Hard Question, assumes the obligation of accounting for the Subject in terms of “a collective dynamic phenomenon that does not require any supervision,” as Dehaene and Naccache put it. This risks seeming to leave out the Subject, precisely because all the work the Subject would presumably have done, once it had enjoyed the show, has already been parceled out to various agencies in the brain, leaving the Subject with nothing to do. We haven’t really solved the problem of consciousness until that Executive is itself broken down into subcomponents that are themselves *clearly* just unconscious underlaborers which themselves work (compete, interfere, dawdle) without supervision. Contrary to appearances, then, those who work on answers to the Hard Question are not leaving consciousness *out*, they are explaining consciousness by leaving it *behind*. That is to say, the only way to explain consciousness is to move beyond consciousness, accounting for the effects

consciousness has when it is achieved.

David Hills:

From early on in your career you've spoken of a design stance it is useful to adopt for various predictive and explanatory purposes. But as Darwinian themes and explanatory strategies became more important in your work, you began talking about "design" as if it were a mysterious je ne sais quois we find less of in some objects, more of in others. You seem to adopt as a regulative hypothesis the idea that nothing can become more designed than it is already, except by either (a) copying into itself some part of the design of a pre-existing thing or (b) benefiting from a design-enhancing generate-and-test procedure, where generation is random and generated novelties are kept only if they pass the test. How does "design" in this quasi-quantitative sense relate to the design stance? How essential is this quasi-quantitative notion of design to the proper appreciation of Darwin's dangerous idea?

I'm glad you asked this question because while I don't think there is anything remotely "mysterious" about design as something "we find less of in some objects, more of in others," I haven't discussed the relation between this and the design stance at any length. As I noted when I first introduced the idea of the design stance, when you adopt it, you get to take advantage of *something* that makes your life easier as a predictor. What? All the design in the designed object. Where did the design come from? It is not a miracle (as Paley or Hume's Cleanthes would claim, in the Argument from Design). The physical work it took to arrange some hunk of the universe in this tidy, predictable way has to have been paid for by something, and *in the end*, the only thing that can pay for design is differential reproduction. In my APA Eastern Division Presidential Address (2000), "In Darwin's Wake, Where am I?" I noted that it's no explanation of the existence of *Hamlet* that it is one of the fruits of the genius of Shakespeare, for where did Shakespeare come from? "What Darwin saw is that design is always both valuable and costly. It does not fall like manna from heaven, but must be accumulated the hard way, by time-consuming, energy-consuming processes of mindless search through 'primeval chaos,' automatically preserving happy accidents when they occur." Design is always normative, always a matter of something that is *supposed to* operate in a certain way to produce certain effects, and when you adopt the design stance, you exploit this by *assuming* that things will work as they are supposed to work. Design thereby always gives you some predictability

“for free;” this happy state of affairs requires explanation, and an evolutionary algorithm of one sort or another is the only thing we know of so far that could explain it in a non question-begging way.

The Dualist:

In “Mechanism and Responsibility,” you claim that the Intentional Stance is a precondition of any moral stance.† You go on to say that “any simple mechanistic explanation of a bit of behavior will disqualify it for plausible Intentional characterization, make it a mere happening and not an action.”† The implication is that any simple mechanistic explanation of a bit of behavior will also disqualify it from being subject to any moral stance.† But we feel that this focus on “simple mechanistic explanations,” as opposed to complex ones, introduces confusion. Where are we to draw the line between “simple” and complex explanations?

We do have to draw lines, separating the moral from the amoral (a major theme in my forthcoming book, *Freedom Evolves*), but life is tough; it won't just give us a sharp boundary on the more or less continuous scale of complexity. And so we watch our children grow into moral agency, moving from a state in which behaviors with simple mechanistic explanations (such as the sucking reflex) get displaced by behaviors—or are they actions?—with more complex mechanistic explanations (such as caressing another toddler who is crying), which soon get displaced by actions whose mechanistic complexities overwhelm us (such as recognizing, and objecting to, unfairness) until eventually we cannot resist including the child in the class of moral agents. We can be fooled (by a child whose apparent solicitude is as utterly amoral as an infant's parent-thrilling laughter) but this should not motivate a search for an essence.

The Dualist:

To what degree do you consider your work interdisciplinary?† Do you feel that cognitive scientists apply your work appropriately?† In your opinion, what would be the ideal relationship between a philosopher and a cognitive scientist?

Almost everything I write has an intended academic audience wider than philosophers and philosophy students. Not everything, since I find myself obliged now and then to engage in some philosophical controversy that I wouldn't expect—or want—anybody aside from

philosophers and their students to attend to. In general I've been gratified by the use cognitive scientists make of my work, and have found that most of my efforts at cross-disciplinary communication have succeeded, but sometimes things backfire embarrassingly. At times like that I like to remind myself of all the enthusiastic misreadings by good scientists of Quine and Popper and Kuhn, for instance. Is the net impact of their labors in the sciences positive? I think so, but the amount of negative to be subtracted from the positive is not negligible. The more you are read, the more you are likely to be misread, I guess. The situation is improving, I think, but too many scientists are either contemptuous of or overawed by philosophers. This is partly because of the well-known tendency of philosophers to hold forth overconfidently on the basis of scant knowledge of the topic, working it all out from "first principles." But at least now we philosophers have company. As I noted in a newspaper interview a few years ago (Sandra Blakeslee, "The Conscious Mind Is Still Baffling to Experts of All Stripes," *New York Times*, April 16, 1996), one of the bright spots at the Tucson II science of consciousness conference was that for once, physicists were rivalling philosophers for the prize of combining arrogance with ignorance of the field. Ideally, I think a philosopher's role in a cognitive science team is a bit like the role of the theoretical physicist vis-à-vis the experimental physicist; you don't have a lab, but you better know just what goes on in the lab, and why. And there are few more bracing tasks than explaining a subtle philosophical point to a smart, impatient cognitive scientist. "Tell me, please, just why I should care about this issue?" Some perfectly fine philosophical issues, worth studying in their own terms, can't pass through such a filter. And you shouldn't try to push them through—one of the only intermittently endearing mistakes philosophers often make. In general, the philosophical issues I care about are those I can get scientists to care about too, but I guess that's just personal taste.

The Dualist:

Universities offer a wide and often confusing range of options to undergraduates interested in studying the mind—from philosophy to cognitive neuroscience to artificial intelligence. What advice would you offer students who want to make progress in this endeavor?

There are so many good paths these days that I am reluctant to discourage any of them, but I will say that I think the recent rebirth of old-fashioned "pure" philosophy of mind is a hopeless dead end. I was appalled to encounter philosophy graduate students recently who

candidly acknowledged to me that one of the main attractions to them of philosophy was that they didn't have to learn anything technical or scientific to do it; they could just live by their wits and read the latest journal articles. I suffered through that when I was a graduate student: ordinary language philosophy flourished and then went extinct, leaving hundreds of oh-so-clever young philosophers with essentially no scholarship, no training, no research program, nothing. (Provoked by these discussions, I have written a little homily on the topic of traps philosophy students can fall into, entitled "Higher Order Truths About Chmess," which is available on my website: <http://ase.tufts.edu/cogstud/papers/chmess.htm>.

The Dualist:

Do you think that scientists do better philosophy or philosophers do better science?

I can think of delightful and appalling examples in both directions. Some scientists' confident "mis-"readings of philosophers have struck me as better value than what the authors intended, and on the other hand, there are some brilliant but pigheaded scientists who perpetuate shoot-from-the-hip philosophical errors that seriously mislead their students, their colleagues, and themselves, and they just won't abandon them. On the other side, philosophers, myself included, have been known to jump to naive interpretations of subtle, complex experimental results and run off half-cocked. But on the other hand, philosophers have also asked some blockbuster questions that have redirected scientific investigations in significant ways.

I guess I find it easier to teach what I consider to be the requisite philosophy to grad students in science than to teach what I consider to be the requisite science to philosophy grad students, but that is mainly a matter of my thinking that one must master rather more material in the latter case. In my experience, a scientist who is both open-minded and a quick study can become quite a savvy philosopher of mind with a few weeks of intensive work (and lots of mid-course corrections from me), but a few semesters of courses are what it takes going the other way. There's just a lot more good cognitive science than there is good philosophy of mind.