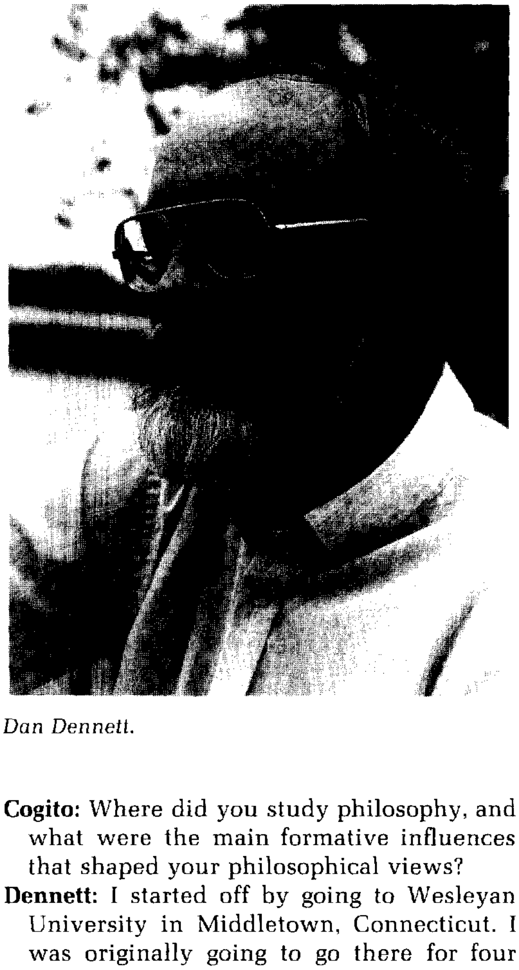
An interview with Dan Dennett

Daniel C. Dennett is Director of the Center for Cognitive Studies at Tufts University, Massachusetts, and a philosopher of mind of international standing. His books include *Content and Consciousness* (1969), *Brainstorms* (1978), *Elbow Room* (1984), and *The Intentional Stance* (1985). He was also joint editor, with Douglas Hofstadter, of the popular collection, *The Mind's I* (1981). His most recent book, *Consciousness Explained* (1991) features in the following interview.



Dan Dennett.

Cogito: Where did you study philosophy, and what were the main formative influences that shaped your philosophical views?

Dennett: I started off by going to Wesleyan University in Middletown, Connecticut. I was originally going to go there for four years. I took a course in logic in my first term. My teachers were under the mistaken impression that I was a budding mathematician, so they put me into a high-powered logic tutorial. Instead of doing a regular logic course, we read Quine's *Mathematical Logic*, and some other books which were much too hard for me. Almost out of despair, one night in the math library, I came across another book by Quine, *From a Logical Point of View*. I read it gratefully, and was fascinated. I thought, 'I'm going to be a philosopher, and go to Harvard and tell this man Quine why he is wrong'. So I transferred to Harvard with it in my head to set Quine straight about various things I thought he was wrong about, as only a freshman could. And in fact I did work on that throughout my undergraduate career. By the time I left Harvard I thought I was one of the great anti-Quineans of the age. I then went to Oxford to do my D.Phil (under the supervision of Gilbert Ryle), and discovered within a few weeks that I was the village Quinean, that I accepted much more of Quine than anyone else in Oxford did. It's clear to me now that both Ryle and Quine had an enormous influence on my work. Maybe I'm what you get when you cross Ryle with Quine and add a little neuroscience.

Cogito: You mentioned Gilbert Ryle, whose famous book, *The Concept of Mind* (1949) was so influential in the 50s and 60s. Yet it is hard to get the students of today to see 'what all the fuss was about'. How much, in your opinion, of Ryle's work is still alive today?

Dennett: One of the facts about philosophy is that it is much more seasonal than students and professors realize. Ryle's work was extremely important at the time in getting people out of a certain slumber they were in.

But he succeeded so well that you don't have to read him any more for that purpose. Now, when we read Ryle, we tend to see the parts he was wrong about, and take the parts he was right about for granted, because everybody thinks that way these days. *The Concept of Mind* was certainly a very exciting book to me when I read it as an undergraduate. That's why I went to Oxford: I thought, 'This is the next step for me — I've got to work with Ryle'. At the same time, I went, once again, in a spirit of controversy, determined to show Ryle what he was wrong about. At first I found him a very frustrating person to work with, because he never fought back. I would try so hard to get him provoked into a disagreement, but he was always so bland and agreeable in response that I thought, 'This is terrible. It's like punching a pillow. I'm just not getting anywhere with this man'. I didn't think I was learning anything from him. Then at the time I submitted my D.Phil dissertation, which was done under his supervision, I happened to go through an early draft and compared 'before' and 'after' to see the difference. Ryle's hand was just all over it — by some strange osmotic process I wasn't even aware of. I wish that Ryle had had more of an influence on the field in one regard, and that is that he was very much opposed to the sort of gamesmanship and one-upmanship that was prevalent in Oxford at the time and is still very much a part of philosophy. I wish that Ryle's spirit of constructive criticism was a little more widespread than it is.

Cogito: Many contemporary philosophers of mind are hostile to the analytic philosophy of the past generation, often dismissing it as a sort of sterile or reactionary scholasticism. You, by contrast, still seem to have considerable respect for the work of Ryle, Anscombe, and others. How do you see the relation between analytic philosophy, on the one hand, and the new empirical work in cognitive psychology and artificial intelligence on the other?

Dennett: When I was a graduate student at Oxford, I was appalled by the way ordinary language philosophy was proceeding in one regard. It was as if you could get a good theory of horses by just studying how ordinary people used the word 'horse', never bothering even to enquire of somebody who might own a horse. That struck me as risky at best. I thought that it might help to see what people knew about horses — or about minds. But what I found when I looked at some empirical psychology was that, once you got used to the technicalities, the very same problems that were addressed at the ordinary language level were addressable there, only *better*, because it was a richer workshop of tools and ideas. There were more data around, but all the interesting philosophical issues had their homes there as well, and some of the methods and techniques of ordinary language philosophy played a good role. So it seemed to me then — and still seems to me today — that ordinary language philosophy taken neat is hopeless; it gives you very modest results of minor (but not quite zero) significance. But if it is coupled with an enquiry into the theoretical presuppositions of the latest empirical work, then you get a much more potent application.

A crude version of the homuncular theory.

Cogito: Until quite recently, accounts of mental processes (perception, say, or memory) that involved a reference — explicit or implicit — to a *homunculus* (a 'little man' sitting in front of an 'inner' video screen; or a little 'librarian' hunting through the archives) were dismissed as *obviously* non-explanatory. You defend (some) homuncular theories against this objection. Can you explain how?

Dennett: The fundamental homunculus objection, which has been around for a long time, foresees an infinite regress. If the little man in your head looking at the little screen is using the full powers of human vision, then we have to look at the even smaller man in *his* head looking at a still smaller screen, and so on *ad infinitum*. That's what's wrong with the little man in the head. However, if you make a simple step, if you break that little man down into a committee of specialists, each of which does less than the full job, has less than the full competence you are trying to explain, now you face the prospect of a finite regress that will bottom out in something purely mechanical. We start with specialist homunculi, no one of which has the full mentality we are trying to explain, and we gradually break these down into simpler functional units, which only by courtesy are called homunculi, and finally we get down to things that are so simple you could replace them with a machine, i.e. with a neurone, or a flip-flop in a computer that only has to remember 0 or 1 as its only expertise.

This idea of avoiding the quandary of an infinite regress with a finite regress seems to me to be actually a much more valuable lesson in philosophy than just this application. The homuncular breakdown is one application, but I think there are others. For instance, consider the idea of an ultimate value. One thinks, 'All values cannot be instrumental, because then we'd have an infinite regress of instrumental values; so there must be a *summum bonum* which is intrinsically valuable'. No: I think there are other ways of looking at the matter which involve the same sort of move. When you reach a phenomenon with a certain sort of complexity, you see that the notion of value doesn't need a foundation in this way. It can peter out in a finite regress without anything that has to count as a *summum bonum*.

One of my favourite examples is the following, which I owe to David Sanford. Consider the following paradox. 'Every mammal has a mammal for its mother. There have not been an infinite number of mammals. Therefore, there have not been any.' The way to break this regress is *not* to find the 'Prime Mammal' from which all others have been born, but to recognize that the boundary between non-mammals and mammals has been gradually passed. The security and reality of today's mammals do not depend on there being a first one, nor on the existence of necessary and sufficient conditions, in the philosopher's sense, for being a mammal. We can let 'mammalhood' trail off indefinitely—but not infinitely—in a finite regress.

Cogito: In your paper, 'Brain Writing and Mind Reading' in *Brainstorms* you attack the notion of a 'language of thought', the idea that our thoughts might be literally inscribed in our brains. It is not wholly clear, though, whether your objections are empirical, *a priori*, or both. If you were to return to this question today, how would you reformulate your arguments?

Dennett: You notice correctly that this argument sits on the cusp between *a priori* and empirical—and for a good reason, because people aren't sure what they mean by 'language'. If they mean something quite literal, then as a matter of empirical fact we know there's no such thing as brain-language. If we then ask, "What must there be?" or "Could there be a system with some of the important, perhaps defining, features of a language, that is involved in the organization of our cognitive lives?" then that becomes many different questions depending on how we cash out the relevant features. One crucial feature that just about everybody hits on is generativity—the fact that an infinity of different contents can be represented very efficiently. That seems to be a feature of human cognition; we, but perhaps no other animals, seem to have a sort of generative cognitive capacity. A question we might ask is, "Does any system, of representation that has generativity count as a language?" Some people might just answer yes, but I think they are missing some possibilities. Consider the following bizarre theory, which as a matter of empirical fact we know to be false. I call it the theory of mental clay. According to this theory, our brains contain a special substance called mental clay, which we use to form little mental clay replicas of all the things in the world that we have a model of. Everyone agrees that we have a model of our world in our brains; this theory just takes that idea literally in supposing that it's a model made of mental clay. So that when grandma dies, you take your mental clay model of grandma, 'kill' it, and put it in a mental clay coffin. It's pretty clear that mental clay is a generative system of representation. But is it a language? If so, if those complexes of mental clay models count as *sentences*, then *anything* surely counts as a language. Now that's a daft theory of mental representation, but there are other possibilities, perhaps at this time still fuzzily conceived, where it's just not clear whether they count as languages. In 'Brain-Writing and Mind-Reading' I was arguing against the over-casual assumption that any system of representation had to be language-like in the way that was assumed. It seemed to me that this involved a failure of imagination on the part of those who were arguing that way.

Cogito: Philosophers of Mind often seem to find themselves debating—ahead of the empirical evidence—what computers can and can't do. Do such debates serve any useful purpose?

Dennett: Well, they sell a lot of books, and permit some people to become famous. Do they serve any other purpose? I suppose they do. Let's look, for instance, at the contributions of Hubert Dreyfus. His book, or manifesto, *What Computers Can't Do*, was roundly criticized by the A.I. community when it came out, and I think it's fair to say that the fate of the ideas in it over the years has been roughly as follows. It is now granted, even by some of his harshest critics, that Dreyfus, although he overstated his case right down the line, identified the hard problems, the really deep difficulties facing A.I. He didn't have the right language, and he misdescribed some of the problems he saw, but he didn't do a bad job of pointing to where the truly troubling issues were. So one might say that he did some good. Or one might say that he didn't really, because they were finding those hard spots anyway and all he did was misdescribe them enough so that he got everybody's blood boiling and created a lot of heat and not very much light. But I would think that, still, history's verdict on Dreyfus is on the positive side, and that his contribution to the debate has been significant. I don't think that anybody else has made a significant contribution with an argument about what computers can and can't do; and I would include in that regard the most recent argument put forward by Roger Penrose, which is simply based on a factual misunderstanding of what A.I. is about.

Cogito: In your paper, 'Cognitive Wheels', you discuss the so-called 'frame problem' facing workers in the field of Artificial Intelligence (A.I.). The problem, roughly speaking, is that of how our mass of *background knowledge* is to be represented, and made accessible to the system as and when required. Opponents of A.I. such as Hubert Dreyfus make a great deal of this objection. Have the defenders of A.I. made any progress towards its solution?

Dennett: I think in fact that the people working in A.I. have made some progress on a number of fronts. There's a new volume entitled *The Robot's Dilemma Revisited*, which is the proceedings of a conference held in Florida a year or so ago where a number of the leaders in the field discussed the frame problem and various approaches to it. My impression from that meeting was that there were a number of very interesting avenues being explored, as well as some dead ends. I think that one of the ways that people have begun thinking about the problem differently—very differently—is by abandoning some of the arch-computationalist ideas (what I call 'High-Church Computationalism') in their setting of the problem. They are taking a more biological approach to thinking about the organization of cognitive systems. I have in mind people like Rodney Brooks at M.I.T. Now you might say that their work is not even approaching the frame problem yet, because they're trying to model simple insects. But at least they are trying to model creatures that can get through their whole lives without falling off most cliffs.

The problem with the frame problem is that it started by assuming that the right underlying model for a cognitive agent acting in the world is the *walking encyclopedia*: you've got a big batch of facts, indexed and organized somehow, and an inference-engine that generates the right inferences from your core set of facts in order to illuminate and guide your current behaviour. Now that's not an inevitable model. There are other models which suppose that there is in fact a division of labour, where you've got one system which is responsible for keeping track of locomotion, for keeping you from bumping into things and looking out, in real time, for the sort of contingencies of that sort which matter; and then you've got another system which can take its time and operate more or less off-line, secure in the knowledge that

An act of proverbial unwisdom.

the 'pilot' is taking care of 'the cliff problem'. Once you have made that division of labour then you can begin to conceive of different ways of making headway on the frame problem.

My own suggestion at that meeting was to take seriously the idea of some anthropologists, that basically, what you have in your head is a whole lot of stories trying to happen. There's 'Three little pigs', and 'Goldilocks', and every folk-tale that you grew up learning, plus lots of things you've seen subsequently. This is the stock-in-trade of simple narratives, which actually have a lot in common from culture to culture. Suppose that the way you're organized is that all of these narrative fragments are sitting there, in parallel, trying to be the truth about the world that you are inhabiting. So you meet somebody new, and one story is saying, 'I bet that's the tin woodman'; another is saying, 'No, it's the big bad wolf', and so on. They are all making tentative assignments of their characters to the people that you meet. This has the effect of getting you to ask roughly the right questions at the right times.

This idea, that you've got all these stories in your heed trying to happen, is actually a nice way of designing a system which will not be at a loss for questions to ask itself. The system will tend to ask the questions that are relevant; but only based on what has proven relevant in the past. We are always susceptible to making an entirely novel blunder. But in fact human beings aren't very good at escaping entirely novel blunders. The main reason we don't paint ourselves into corners more often is that we have heard—and remembered—stories about people painting themselves into corners!

Cogito: In 'Why You Can't Make A Computer That Feels Pain' (in *Brainstorms*) you set out to write a 'Pain-Program'. The exercise reveals that what seems, at the phenomenological level, to be a single simple *quale*, may turn out, at the sub-personal level, to be an extremely complex functional state. What does this reveal about the amount and quality of the self-knowledge that we can obtain through introspection?

Dennett: I think that introspection is a remarkably unreliable source, that our access to our own mental states is not only in most regards not privileged; it is also, as Keith Gunderson once marvellously said, actually under-privileged. We are designed to have pretty good ideas about what's going on in us, but they tend to be couched in metaphors and subject to systematic distortion, so that our authority is limited and in the end vanishing. Wittgenstein once gave the example of a man who says, "I know how tall I am—I'm *this* tall", placing his hand on the top of his head. Well, there's a person who is not wrong, but only because he's not right about very much. He has reduced his claim to a logical minimum. And if we do this we can arrive at something that no one would be able to show to be wrong, but such claims have lost all interest.

Cogito: At the heart of your philosophy of mind is the notion of an *Intentional System*, a system (animal, robot, human, or whatever) to which we *attribute* beliefs and desires as part of an explanatory theory by which we attempt to understand the system's behaviour. This theory has laid you open to accusations of two sorts of anti-realism: (1) that there is no *real* (i.e. objective) distinction between systems that are, and those that are not, Intentional Systems; and (2) that even for a system that we have agreed to regard as an Intentional System, there may be no *right* answer to the question of what it 'really' believes or desires. How do you answer these objections?

Dennett: The first of these objections is one I think I've already addressed in talking about the breakdown of homunculi. The question, 'At what point do we stop having *real* intentionality and have mere *as if* intentionality?' is like the question, 'At what point do proto-mammals give way to mammals?' The question doesn't have to have a principled answer for it to be true that human beings, and some of their sub-parts, are real intentional systems.

Now what about the other question, as to whether or not there's a fact of the matter? Here I claim that, since at the heart of any system of intentional attribution there is an idealizing assumption of rationality —ultimately a normative assumption—it follows that in those cases where there is a falling short from optimal rationality, there's just no fact of the matter at all about what the truth is. I don't view this as a surprising, or metaphysically extravagant or unsettling result. It seems to me to fall into a rather ordinary variety of facts—where the conditions that are presupposed for the proper application of some term lapse, there is no fact of the matter about what the proper application of the term is.

Cogito: In regarding another person as an Intentional System, i.e. ascribing to him or her beliefs and desires, we are committed, in your view, to interpreting most of those beliefs as *true* (or at last rational). Many students find this claim quite incredible. Have you found any way of convincing such doubters?

Dennett: I've found a way that convinces *some* of them. I don't know what I can do about the residual doubters, except throw up my hands and wish them well. I think that the source of this doubt is their acceptance of a familiar philosophical bit of misdirection. They think of beliefs as sentences deemed true; and moreover they think of beliefs as a rather special sort of these, i.e. the ones considered, reflected on, and asserted. In fact, these are the states I would call *opinions*, and they represent a vanishingly small proportion of our beliefs. They are not themselves, I want to say, beliefs; they are, as it were, the offspring of beliefs. Right now, you have extraordinarily many beliefs about the room you're in and about what's happening in front of you. These are beliefs that are produced by your eyes, ears, and other sense organs. You have lots and lots of beliefs about where you were born, what you have done today, and so on. When we are talking about the majority of your beliefs being true we have to include all those. Your beliefs about physics, about philosophy, about the nature of truth—your theoretical beliefs—are not even close to being one millionth of one per cent of the information that you have and rely on. Now when students and other sceptics say, "I can imagine a creature most of whose beliefs are false", I just don't think they can. They haven't done the job of imagining right. They can't coherently describe an agent, most of whose beliefs are false, that can be treated as a believer at all.

There's a certain cartoon character who comes close — the nearsighted Mr Magoo. Mr Magoo informs us by his almost continuous monologue of how he takes the world to be, and he's always dead wrong, and at great risk because he's misinterpreting everything because of his colossal nearsightedness. But of course the joke is that it always works out; that is, the crocodile never quite closes its jaw on Mr Magoo; and he ends up back safe and sound in his own bed at the end of the day. It's a wonderful comic conceit, and I think it proves my point. One recognizes that the world *cannot* be this way, not for more than a few minutes. The nearsighted Mr Magoo is impossible for a fairly deep reason. You can't survive to adulthood unless most of your beliefs are true.

Cogito: Well, that both leads us on to our next question and, I think, ties up neatly with what has gone before, namely your attack on Cartesian assumptions in general. It is surely easier for me to think of the possibility of most of my beliefs being false if I think of myself as a sort of disembodied spirit, than if I think of myself as a biological organism and a product of evolution by natural selection. Now in your later book, *The Intentional Stance* (1987) you seek to 'ground' your philosophy of mind in a philosophy of biology. How much support do you think Darwinism gives to your particular views about the mind?

Dennett: I haven't a settled opinion about this, because my estimate keeps growing. That is to say, the extent to which I think the theory of evolution by natural selection is necessarily implicated in any sound philosophy of mind keeps growing. In a recent article called 'The Interpretation of Texts, People, and Other Artifacts' [in *Philosophy and Phenomenological Research*, Supplement, 1990] I make the claim that interpretation, whether it is of texts, as in hermeneutics or literary criticism; or of people, as in folk-psychology; or of artifacts in general, is always the same game — that is, there is one exercise that's going on, and it's grounded, all of it, in evolutionary considerations. The principles of reverse engineering that one must apply in order to make sense of minds, or of their products, turn out to be the very same principles that are used by biologists to make sense of organisms, and ultimately it's all grounded in the theory of natural selection. That's why my next book is going to be on Darwin — it's called *Darwin's Dangerous Idea*.

Cogito: In your current book, *Consciousness Explained* [In the USA, Little, Brown, & Co, 1991; in the UK, Allen Lane, 1992] you propose a 'Multiple Drafts Model' of consciousness. Your theory is extremely subtle and complex. It is probably unfair to ask you to summarize such a rich book in an interview, but perhaps you could give us a rough idea of the key features of the theory?

Dennett: I would say that my view involves a family of theories, rather than a single theory. The alternative is the view that I call 'Cartesian materialism', or 'the Cartesian Theatre'. The Cartesian materialist says, "Everything Descartes said about the mind was true, *except* of course the claim that it's immaterial. The mind is just the brain, but there has to be a material place in the brain where the contents of consciousness are presented". To whom? To the inner witness. We seem to be right back where we started from, with a presentation and appreciation process happening somewhere in some special inner sanctum in the brain. Now we know there's no place in the brain where it all comes together. Yet we find it very hard to conceive of how consciousness could exist if that were not the case.

I'm trying to sketch, in effect, a family of alternative theories. That's why it has to be, in some regards, very metaphorical, because I want to be neutral with regard to some possibilities I am in no position yet to resolve. The main point of my alternative is that the appreciation, the 'taking' of the 'given', has to be broken down, fragmented, distributed around in space and time in the brain. This has implications that are far from obvious; indeed, some of them are profoundly counterintuitive. My task in *Consciousness Explained* is to make the theoretical world safe for these alternatives, so that they no longer seem so profoundly counterintuitive, so that we can begin to think positively about their details.

Cogito: One of the main misconceptions that you attack is the idea that there is a single point or centre in the brain which is the seat of consciousness, the 'place where it all comes together'. You go on to argue: "Since cognition and control—and hence consciousness—is distributed around in the brain, no moment can count as the precise moment at which each conscious event happens (p 169)." But surely the brain-events which underpin conscious experiences must occur at definite times, even if they occupy different places in the brain?

Dennett: Sure. The individual brain-events are clockable down to the millisecond, if you have the right equipment, but my claim is a conceptual point about what does not and cannot follow from that. There is a nice analogy. Let's consider the British Empire, which was distributed all over the world in the nineteenth century. Now let's consider the notorious case of the battle of New Orleans, which was fought fifteen days *after* the truce was signed in Belgium, ending the war of 1812. The battle of New Orleans was an intentional action of the British Empire. So was the signing of the truce. Yet they happened in an unfortunate and anomalous order. Now suppose somebody asks, "When did the British Empire learn of the signing of the truce?". Suppose we can tell when the ambassador in Belgium knew it, or the commander of the British forces in New Orleans, or Parliament, or the monarch, or the Governor-General in Calcutta. Each of *those* questions can be answered to the second, or at least to the day. But if you ask, "When did the British Empire learn of the truce?", the question is ill-formed. We can reply, "Oh, sometime in early 1815", but that's as close as you can get.

Now compare that question to this one. The subject asks, "When did I become aware of the red disc turning green in Kolers' phi-phenomenon? Don't tell me what happened on my retina; don't tell me what happened in my vocal apparatus at the output end; don't tell me what happened in area V1 in my brain; I want to know when *I* became conscious of it". Same answer. *You* are not located in any one place in your brain. So, just like the British Empire, the question of when you first learned this fact has no determinate answer.

It's interesting that when I ask the first question, about the British Empire, a lot of people initially think, 'Well, we'll have to find out when *the King* learned'. Fine. Maybe a case can be made for: 'the Empire, *c'est moi*', in the case of the king. But there isn't any king-homunculus' in your brain; there just isn't any one place in the brain where the message has to get through so that that's when I learn about it. So this idea that there isn't any way of determining, to the millisecond, when *a person* is conscious of something, is simply the same conceptual point at a different time-scale.

Cogito: May we pursue the relation between time and consciousness a bit further? In chapter six, which begins with a quotation from Kant, you make several Kantian-sounding points. For example, you say (p 148) that the representing by the brain of *A before B* does not have to be accomplished by, first, a representing of A, and then, a representing of B. You point out that a person can judge, in a single momentary act, a proposition about a temporal sequence. Suppose, however, that we consider, not a person judging that A occurred before B, but rather a person actually *experiencing*, in real time, a process in which A occurs then B occurs. Doesn't this person *have* to experience A first, and then experience B?

Dennett: The way you have worded the question, the answer may well be yes. But consider a slightly different case. Consider a man experiencing something illusory. There isn't an A first, followed by a B; but nevertheless there is an experience of A-followed-by-B. Now here you're suggesting a distinction between experiencing and judging that I want to combat. Let's take a very simple case of perceived motion. You see an arrow move where no arrow moves—there are just two separate flashes of light. Now it's tempting to think that for you to see the arrow move, if the arrow out there doesn't move, then the brain has to make an arrow, make it move, and let you look at it. But that's a fallacy. You can see the arrow move without the brain having to construct a moving arrow for you to look at. It can be the case that you see the arrow move even though the arrow 'out there' doesn't move, and no arrow 'in here' moves either.

I never know whether that point is just obvious, or whether it's really counterintuitive, but I think it's a truth that can be arrived at by many different paths.

For example, V. S. Ramachandran has a wonderful experiment in which subjects look at a TV set on which there is just snow, twinkling little spots of light randomly distributed. You now draw a little X in the centre and ask the subject to fixate on X. To the side of the X, you then fix a little piece of grey cardboard. If the subject continues to fixate on X, the grey cardboard will gradually be filled in with twinkle and will disappear. Then, marvel of marvels, if you arrange suddenly for the background to become uniform (without twinkle), there will be, for a brief period of time, some twinkle left over in the square. A stunning and surprising experimental result. This is twinkling that isn't in the world, and never was in the world. Here is the relevant philosophical question. Can the brain represent twinkling without representing lots of individual twinkles? Does the brain's representation of twinkling *have* to consist in its representation of a lot of individual twinkles? The answer is no. The brain does not have to represent all the individual twinkles for you to *perceive* twinkling. Perception does not involve the presentation of a display in a Cartesian Theatre.

Cogito: You claim that "there are no fixed facts about the stream of consciousness independent of particular probes". Scientists in many fields encounter the practical difficulty that their attempts to measure something inside a system *interfere* with the system, and consequently affect the properties of the thing they are trying to measure. So they never obtain data that are interference-free. But you seem to be making a much stronger claim here. Could you please elucidate further?

Dennett: I am making a stronger claim, because the phenomenon we are talking about is a phenomenon which is, if you like, by definition a matter of probing. That is to say, to talk about a conscious phenomenon that exists unperceived or unobserved is to talk apparent nonsense. Curiously enough, some attempts to account for consciousness have the feature of trying to establish the category of the objectively subjective, which is, I think, just a covert contradiction. There is no such thing. Here is the one place where verificationism reigns, and obviously *should* reign. The astronomers' rule of thumb, 'If you didn't write it down, it didn't happen', nicely overstates the verificationist thesis, but when it comes to consciousness, that's just the truth. 'If you didn't experience it, it didn't happen (in experience)'—that's just a tautology. But to say that you experienced it either means that in some sense you 'wrote it down', that it was a probe that led to further effects, or it doesn't mean anything.

This is particularly clear, I think in the case of meta-contrast, or backward masking. If you are shown a flashing red disc on a screen, you have no problem at all in seeing it and in counting the flashes. But if, very briefly after the disc flashes on, a 'doughnut' shape is flashed, so that the disc would have been, as it were, in the 'hole' of the doughnut, then you get a sequence of flashes, disc-doughnut, disc-doughnut, disc-doughnut, etc. However, all that you will then see—all that you will report seeing—is the doughnut. It is as if the disc were not there at all. How do we explain this fascinating result? There seem, at first sight, to be two possibilities. The standard explanation in the literature is what we might call the ambush theory. This says that the disc-stimulus is on its way up to consciousness; when somehow the doughnut-stimulus overtakes it and ambushes it on its way, so that all that ever gets into the door of consciousness is the doughnut. But notice that there might be another explanation, i.e. that you do become aware of the disc in consciousness, but that immediately afterwards the doughnut rushes through the gate of consciousness and not only impinges itself on your memory but also erases your memory of the disc having been there a moment before.

Now we seem to have two distinct theories. One of them says that there's a post-experiential tampering with memory by a sort of Orwellian Ministry of Truth. The other says that there's a pre-experiential expunging of the stimulus on its way up to the Theatre of consciousness. I argue that this is an illusory disagreement. There can't be any fact of the matter as to which side of the Theatre the interference occurs, because there isn't any Theatre. Manifestly, the second stimulus interferes with the processing of the first stimulus, but if we ask ourselves whether this is post-experiential memory processing or pre-experiential presentation processing, there just is no answer to that question.

Cogito: Your colleague Douglas Hofstadter remarks, on the dust-jacket of the book, 'While *Consciousness Explained* is certainly not the ultimate explanation of consciousness, I believe it will long be remembered as a major step along the way to unraveling its mystery'. Do you agree with the first bit? And if so, what, in your view, are some of the mysteries that still remain?

Dennett: Oh, I agree with the first bit in this sense. I think that the hard empirical work remains to be done. I said before that my theory is a sketch. It is deliberately neutral with regard to some big puzzles. But they are puzzles, not mysteries. The difference is that with a mystery you don't know how to proceed, whereas with a puzzle you have a good idea about what the ground rules are and about what constitutes progress. In my braver moments I think I've turned the mysteries of consciousness into a series of puzzles. But that still leaves a lot of work to do.

Cogito: So turning mysteries into puzzles is a legitimate aspiration for the philosopher in general, and the philosopher of mind in particular?

Dennett: I think so, yes. In fact, I think that's an ancient and well-regarded tradition in philosophy, the idea that philosophy starts with the mysteries, and when it can turn a mystery into a puzzle it gives birth to a new science. Of course we already had the sciences of psychology and neuroscience, but they were, in a sense, prematurely born, so there's still work for the philosopher to do before they can simply proceed with their puzzles.

Cogito: Perhaps we could end this interview with one or two questions of a less technical nature. Within psychology, there has long been a tension between so-called 'scientific' and 'humanistic' approaches to the subject. Scientific psychologists have often argued that humanistic psychology just perpetuates confusion, falsehood, and wishful thinking. The challenge may come from Skinner's behaviourism, or from the eliminative materialism of Paul and Patricia Churchland, but the message is that our 'folk-psychology' is under threat from the advance of science. How do you regard this debate? Are scientific understanding and human dignity compatible?

Dennett: I think there are several issues wrapped up in your question. The question of whether folk-psychology has a future can be split into two. One might ask, 'Does it have a role in scientific psychology?', which is really what the debate between humanistic psychologists and Skinnerians was about. There, I think, the answer is that it does, but not the role that someone might think. It's analogous to setting the specifications for an engineering problem, rather than being a description of the solution. We use folk-psychology to partially characterize the nature of a competence to be implemented; and once the implementation has been created, we may see that the traditional folk-psychological way of talking about it no longer has application.That's not a negligible role at all for folk-psychology within scientific psychology—in fact, I think it's an exalted and very important role.

But that's quite distinct from the question of whether, as materialistic neuroscience marches on, there is going to be room for our notion of human dignity. I think the answer there is manifestly *yes*, but with different foundations. Anybody who supposes that our vision of ourselves as moral agents, or as lovable, or as having dignity, depends on a certain dualistic conception of ourselves as harbouring a special sort of spirit, seems to me to hold a remarkably trivial and trivializing notion of the source of human dignity. It's as if we each carried a special gold coin without which we were trash—a stupid idea when you look at it that way. Our appropriateness as objects to love, respect, admire, or care for, is much more a function of our complexity, and of the kind of complexity that we manage to have. If anything, the account of our worthiness that one gets from a mature cognitive science is vastly preferable to the one that you get from the traditional vision. And this is not just a theoretical issue—it's also a practical one. Ethical issues regarding abortion, or prolonging the lives of brain-damaged patients, for instance, require informed moral decisions. At the moment I think that the traditional moral categories are just getting in the way. They are not doing any work, and sometimes they are positively harmful. I do not believe that progress is made on the abortion debate by asking, 'At what point does the soul enter the foetus?' I think that is a manifestly preposterous way to engage in the debate. Well, then, let's find some other and better terms.

Cogito: Do you see your technical work in the philosophy of mind as helping to solve such issues as those mentioned above? Might it, for example, help us to formulate a credible *moral psychology*?

Dennett: In fact, I have long aspired to extending and developing my ideas in this direction. *Elbow Room* fills part of the gap, as do the papers, 'Conditions of Personhood' and 'Mechanism and Responsibility'. I think there is a natural extension of my ideas into the area of moral psychology, and I am delighted to say that a wonderful inroad has been made by a colleague of mine, Stephen White, in his new book, *The Unity of the Self*. Although his theory of mind, in that book, is almost entirely consonant with mine, the book is far more than just an extension of some of my ideas—it also goes deep into areas which I hadn't imagined at all. The book seems to me to go a long way to demonstrate how to tie ethics and philosophy of mind together in the field of moral psychology.

Cogito: I hope you will forgive us if we end with a more personal question. Do you find that your work in the philosophy of mind helps you to gain a greater understanding of yourself and others? Are you a better husband, father, or grandfather as a result of your work? Can you think of examples where your theoretical work has guided your personal life, or *vice versa*?

Dennett: Actually, when I think about this I realize that the task of parenting, for instance, has got to go on in real time, and under such tremendous emotional pressures that it would be very unlikely for anybody's theory to play more than a passing role. There's very little evidence of a positive correlation between anybody's psychological or philosophical theories and their capacity as a parent. Not because their theories are good or bad, but just because parenting is such a difficult task.

I can think of some trivial—or perhaps not so trivial—cases from my own experience. One that I'm rather proud of occurred when my daughter was a little girl of about five. She was doing acrobatics on the piano stool, and it fell and crushed some of her fingers. She was not only in great pain but terrified, almost hysterical. I realized I had to do something to calm her down. I'd just finished my research on pain (for 'Why You Can't Make A Computer That Feels Pain'), and had an idea that I thought might work. So I tried it. I said to her, holding her crushed fingers close to the palm of my hand, "Andrea, here's a secret. Push the pain into my hand! Quick! You can do it!" And immediately she stopped crying, and a smile formed on her face. She had 'pushed the pain into my hand'. This was a sort of impromptu hypnotic suggestion on my part that worked remarkably well. I told her, "If the pain starts coming back, push it back into my hand again", which she did. Later, we left the pain on a wall. "We'll leave it there", I told her, "and go to the hospital". She was quite content to accept the theory. In her distraught state she wasn't being very critical. But it does show that sometimes all it takes is a new theory to alter your consciousness quite dramatically. This seems to me to be a pretty good instance of it.