Dennett's Position in the Intellectual World

by Andrew Brook and Don Ross

1. Dennett's Life and Work

Over the past thirty years, Daniel Clement Dennett has had a major influence on our understanding of human intentionality and agency, consciousness (and thereby phenomenology and the architecture and neuroscience of consciousness), developmental psychology, cognitive ethology, artificial intelligence and evolutionary theory. In this introductory essay, we will first give a chronological survey of these contributions and then, starting with Dennett's place in the intellectual history of the last half of the twentieth century, construct an overview of his philosophy. Dennett has played a central role in one of the most significant theoretical revolutions of the past fifty years, the cognitivist revolution. This revolution demolished simple empiricism and put in its stead a view of human action as requiring interpretation in terms of a rich reservoir of cognitive resources and, many argue, evolutionary history. Dennett has played a role in this revolution for thirty-five years now.

Recently, a number of collections have appeared on Dennett's work. This one is different. We are interested in the influence Dennett has had beyond the bounds of academic philosophy. To assess this influence, we have assembled a team of experts who either specialize in one of the areas in which Dennett has had such an influence (Simon Baron-Cohen in developmental psychology, Robert Seyfarth and Dorothy Cheney in cognitive ethology, Yorick Wilks in artificial intelligence) or who, though trained in philosophy, have attained expertise in a discipline beyond philosophy (in alphabetical order: Akins in neuroscience, Brook in cognitive science, Clark in cognitive science, Churchland in neuroscience, Ross in economics and evolutionary theory). We wrote the introduction as much for these contributors as for our readers. Rather than each contributor trying to summarize the aspect of Dennett's work most relevant to his or her contribution, we put together a single common overview for all to share. The various contributors then took this overview as read when they prepared their essays. Thus the essays should be read together with this introduction.

Though Dennett has deep roots in American pragmatism, he was actually born in Lebanon (during WWII; his father was representing the US government there) and he did his doctorate at Oxford. He is a seasoned world traveller and has spent time in most of the major universities of the western world (including, he once said with some pride, every university with a philosophy graduate programme in Canada). He did his undergraduate work at Harvard, where, he says, he vigorously resisted his teacher and the most influential American philosopher of the 20th century, Willard van Orman Quine (doctrines for which he later developed more sympathy) (Dennett 1998, p. 357). His doctoral studies at Oxford were done under the tutelage of the most influential Oxford philosopher of his time, Gilbert Ryle. Dennett received a D. Phil. in the remarkably speedy time of two years in 1965.

Dennett's life has been as stable and unsensational as his work has been brilliant and influential. After a brief stint at the University of California Riverside, he has been at one university ever since, Tufts University in Medford, Massachusetts, a suburb of Boston. He married in university and has been married to the same woman ever since. They have two grown children. He runs a working farm in Maine. (Dennett is said to get some of his best ideas while working the land. Douglas Hofstadter calls this 'tillosophy'.) He is an expert high-seas sailor and navigator and an accomplished drummer and choral singer. He is one of a very few philosophers who commands bestseller-sized advances for his books. Dennett once described himself as an ACLU (American Civil Liberties Union) liberal.

Dennett is famous for his generosity to students. A small army of increasingly important young philosophers, psychologists and even MDs have passed through his Centre for Cognitive Studies at Tufts as post-doctoral fellows or visitors. Though Dennett has had a deep influence on thousands of researchers, there are few ‘Dennettians’. Unlike many world-class intellectuals, Dennett has never sought to create disciples. His former students are remarkably independent-minded.

Dennett is an accomplished philosophical humourist. He was the main force behind *The Philosopher’s Lexicon* (1987d), which consists of ‘definitions’ of philosopher’s names that reflect (or at least spoof) some aspect of their work. The entry for ‘Dennett’ (not written by Dennett) reads in part: “Dennett, v. To while away the hours defining surnames; hence, *dennettation,* n. The meaning of a surname. ‘Every surname has both a meinong and a dennettation’.”

There are many, many sides to Dennett’s contributions but one of his most important contributions has been to challenge unexamined orthodoxies. One of his characteristic ways of doing so is to go after comfortable assumptions with what he calls intuition pumps. Here is an example. Prereflectively, most of us would think that there is a clear difference between how something tastes to us and how we react to that taste (with pleasure, indifference, disgust, etc.) But consider the case of Mr. Chase and Mr. Sanborn:

> Mr. Chase and Mr. Sanborn both used to like a certain coffee. More recently it has lost its appeal. The reasons they give seem to differ markedly. Chase: ‘The flavour of the coffee hasn’t changed but I just don’t like that flavour very much now.’ Sanborn: ‘No, no, you are quite wrong. I would still like *that* flavour as much as ever. The problem is that the coffee *doesn’t* taste that way any more.’ [taken from Dennett 1988a]

We are meant to say to ourselves, ‘Hmmm, maybe this distinction I want to draw between how something tastes and how much one likes the taste does not correspond to a real difference.’ Then we are meant to generalize the doubt: ‘Well, if there isn’t a clear boundary to be drawn here, what about with other mental states?’ – And we are well on our way to shaking up our traditional philosophical conception of the mind as a place populated by a bunch of clearly demarcated mental states.

Dennett is not just a philosophical gadfly. The unsettling intuition pumps, awkward rhetorical questions, etc., are used not just in the service of iconoclasm (though they are used in the service of iconoclasm). Dennett has a deep philosophical mission, one articulated in his very first book and carried through with verve, ingenuity and great continuity every since.

His first book grew out of the work he did with Ryle. It was called *Content and Consciousness* (1969). These two words, 'content' and 'consciousness', encapsulate much of Dennett's mission. 'Content' refers to the contents of the mind: all the beliefs and desires and values and emotions and hopes and expectations and memories and ... and ... and ... that make up the mental life of all cognitively intact human beings. And 'consciousness' refers, of course, to our consciousness of our world, our mental contents, and ourselves.

In Dennett's view, the correct order in which to examine these topics is content first, then consciousness. (The essays in the current volume might seem to be in the reverse order but they are not. We will explain why later.) To oversimplify mightily, he devoted the 1970s and 80s to content (with some important forays into consciousness) and the 90s to consciousness (with some important additional work on content). Mentioning just these two topics may make Dennett's contribution look narrow. It is not. To the contrary, his work on consciousness has led him to study how consciousness evolved, pathologies of consciousness such as Dissociative Identity Disorder (what used to be called Multiple Personality Disorder), whether there is any real difference between how a mental state functions in us and how it feels to us (what the philosophers call, somewhat quaintly, its *qualia* or felt quality), what "selves" might be, methods for studying consciousness, how to model consciousness as a cognitive system, the nature of introspection (the consciousness we have of ourselves and our own mental states), the neural implementation of consciousness, and so on – just about every issue in connection with consciousness that one could think of. And his influence has reached to just about everybody interested in any of these issues.

Similarly with content. His work on mental content has led him to questions about artificial content (AI), the evolution of content, the relationship of content to the environment and brain (neuroscience), content in nonhumans (cognitive ethology), the nature of explanation in psychology and science generally, how content is represented and the different styles of mental representation, the relationship of representations to the brain, how we ascribe mental content to ourselves and others, and so on – all the issues alive in current work on the mind, how it evolved, and its place in the world.

His first major work, as we said, was *Content and Consciousness* (1969). It will figure prominently in the second part of this introduction. His next book was a collection of essays written during the '70s, *Brainstorms.* This work not only brought together an extraordinarily interesting group of papers on mental content (and four on aspects of consciousness), it helped launch a unique publishing enterprise, Bradford Books. Founded by Harry and Betty Stanton and subsequently absorbed by MIT Press, the Bradford Books insignia has become one of the most important collections of books in philosophy of mind and cognitive science in the English language. Except for some trade books of the '90s, all Dennett's books since *Brainstorms* have been published under the Bradford insignia (*Brainstorms* has recently been rereleased as a Penguin softcover).

Brainstorms begins with the first full articulation of Dennett's distinctive approach to mental content. The approach is called the intentional stance and the paper is called 'Intentional Systems'. Says Dennett, we can approach something in order to explain it from three stances, the physical stance, the design stance, and the intentional stance. Each has its own advantages and costs, as we will see when we examine some of the details of the approach in section 2.2. This mode of explanation yields impressive results when we seek to understand people's reasons for what they say and do and in other areas. The best current theory of autism, for example, is that autistic people lack the implicit notion that others are intentional systems, so their ability to adopt this stance to themselves and others is impaired or absent (**Griffin and Baron-Cohen**). (When a paper in this volume discusses a topic, we will so indicate by putting the author's name in bold.) The approach also yields impressive results in cognitive ethology (**Seyfarth and Cheney**). Although few economists are aware of the connection, the game-theoretic logic that underlies the intentional stance has also become one of the main approaches in economics (**Ross 1)**.

As well as these system-anchoring reflections on content, *Brainstorms* also contained a number of papers on behaviourism and its early cognitivist replacements, a fascinating paper on AI as what you might call philosophy under the discipline of reality (you have to implement your ideas in AI) (**Wilks)**, no less than four papers on aspects of consciousness that presaged a book still fifteen years away, and a number of papers on decision-making and responsibility.

It was decision-making and responsibility that Dennett turned to next, in an idiosyncratic little book called *Elbow Room: The Varieties of Free Will Worth Wanting* (1984). Beginning life as John Locke Lectures in Oxford, the book advocates a brisk compatibilism between decisions being causally determined and decisions being free in any way that is "worth wanting". For better or for worse, this book has had less influence than Dennett's other books. Interestingly, he is working on a new book on free will as we write (in the middle of 2001). People are awaiting his new thoughts on the matter with great anticipation. (Strictly speaking, *Elbow Room* was not his next book after *Brainstorms. The Mind's I* edited with Douglas Hofstadter (1981) came in between. It is mainly a collection of works by others, however, so we won't comment on it here, quirky and entertaining though it is.)

The year 1987 saw his second major collection of papers on content, *The Intentional Stance.* Dennett tells us that the first paper in the collection, 'True Believers', replaces 'Intentional Systems' as his flagship paper. The main difference between the two is that in determining what beliefs, desires, etc., to attribute to a system, i.e., what beliefs, desires, etc., the system should have, evolutionary considerations now play a much bigger role than they did in the earlier paper (**Ross 1**). The concern to get clear about the "ontological status" of mental states that was finally put to rest in ‘Real Patterns’ is very much in evidence in Dennett’s reflections in the early part of the collection. Two later papers in the collection are on evolutionary theory, including one of the most famous pieces he has ever written, ‘Evolution, Error, and Intentionality’, a paper in which he explicitly and very firmly sets his face against all forms of the idea that mental content can ever be intrinsic to brain states. The volume contains the first full expression of the method for studying consciousness that was to become the centre of a large book on consciousness four years later, in a paper called (enigmatically) ‘Beyond Belief’. There is nothing else in the collection on consciousness. This is firmly a volume about content.

At this point Dennett left content behind for a while and turned to consciousness. His next book was a huge, sprawling work called *Consciousness Explained (CE)* (1991a) (For Dennett, modesty is a virtue to be kept for special occasions.) *CE* was aimed at a wide audience. For the first time (with the exception of *The Mind’s I),* Dennett chose a trade publisher. It would not be the last.

In *CE,* Dennett has two main targets. One is the picture of conscious states that the tradition received from Descartes. This is the idea that there is something to a conscious state, some felt quality, that is unmistakeably clear and clearly different from all other properties of mental states (**Brook**). The other is the picture of the conscious system that the tradition received from Descartes. This is the picture of the conscious system as a kind of screen on which conscious states play before a little homunculus sitting in the middle of the theatre (Dennett calls it a Cartesian theatre), where the conscious states themselves are conceived of as discrete, separately identifiable states, states with, for example, clear stop and start points. Dennett wants to replace the Cartesian picture with what he calls a Multiple Drafts Model (MDM) of consciousness. The MDM treats consciousness as a kind of mental contents, almost a matter of programming, which is controversial (**Churchland**). The book concludes with a chapter pulling together the picture of the self that we’ve already seen and a final attempt to beat back two of the more esoteric attempts to make consciousness mysterious, one each by Thomas Nagel and John Searle.

Both Dennett's theory of content and his theory of consciousness require that the brain have certain capacities and structures. It must have the capacity to produce that incredible array of behaviour expressing mental content that we find in ourselves and others. And it must house a MD-type structure, which, Dennett suggests, probably consists in a Pandemonium architecture of some kind (**Akins**).

Having settled accounts with consciousness, Dennett next took up a task that many had been expecting him to turn to for a long time, evolutionary theory. *Darwin's Dangerous Idea (DDI)* (1995) was also published as a trade book and enjoyed the same phenomenal success as *CE.* Two best sellers on abstract issues in the philosophy of mind and philosophy of biology in a row is not bad!

In *DDI*, Dennett argues for two main claims:

1. Darwin's theory of evolution is a powerful 'universal acid' for dissolving all manner of intellectual 'skyhooks' and other pseudoscientific props that philosophers (and not just philosophers) have wheeled onto the stage to try to patch up hopeless theories;

and yet,

2. Darwin's theory of evolution may deflate the pretensions of many accounts of morality but the ones it deflates are highly problematic in any case. Contrary to those who see Darwin as the destroyer of all morality, however, the theory of evolution leaves one perfectly satisfactory approach to morality and political philosophy untouched. namely, traditional western liberalism.

In the course of developing his picture of evolution, Dennett tackles the central debates in contemporary evolutionary theory: adaptationism vs. the idea that much of evolution has consisted in good tricks developed for one function being co-oped for other functions; smooth evolution vs. Gould's punctuated equilibrium; the role of genetic drift, climate changes, and other such accidental elements in evolutionary change; and so on. As readers of the *New York Review of Books* will know, some of Dennett's claims on these issues generated a firestorm of controversy. Though it's a little hard to understand why, Dennett found himself accused of being an ‘ultra-Darwinian’ (whatever that is) and Dawkins’ lapdog. Needless to say, he responded with equal vigour and a hot argument ensued (**Ross 2**).

Among the most important claims introduced in *DDI* is a claim that it is language that makes it possible for us to have our kind of mind, a kind of mind that, by being able to cooperate with other minds and record the results of cooperation for others to build on, can figure out the physics of the universe, find cures for most serious diseases, build Hubble telescopes and Channel tunnels, create artificial speech-interpreters and problem-solvers, and so on and so forth. By endowing us with language, evolution has utterly separated us from all cognitive systems that do not have language, where by “language” we mean something that is syntactically-articulated and functions by building sentences that are structured compositions of lexical units. This became the basis of his next book, *Kinds of Minds* (1996), an attempt to pull some of his most important ideas about minds in general and our kind of mind in particular together in one place and to say more than was said in *CE* (or anywhere else) on the evolution of our kind of intentionality. How far Dennett wants to go here, whether for example he wants to restrict even consciousness to animals with our kind of language, is not entirely clear (**Clark**).

Dennett’s most recent work (as of the time of writing) is *Brainchildren* (1998), a collection that pulls together a remarkably diverse array of pieces written over the previous decade or so and appearing in various sometimes quite obscure places. The pieces range widely. They include rich philosophical essays such as ‘Real Patterns’ (already mentioned); responses to criticism, especially of *CE;* a strikingly wide group of papers on artificial intelligence and artificial life; and some new papers on animal cognition and consciousness. The collection closes with two occasional pieces, a self-portrait and one of Dennett’s few forays into morality, a paper on what the dangers of information technology are and are not. This collection is fun to read – much of it is written in an even more relaxed, polemical style than is usual with Dennett. One of the papers on animal cognition is on animal pain. For interpretationists about mental life such as Dennett, pain poses a special challenge because if anything in mental life just is, just hits us no matter how we interpret it, pain would seem to be it. Dennett does an impressive job of ‘downsizing’ the range of pains that animals could plausibly be said to feel.

In addition to the books we have discussed, Dennett has written hundreds of papers, critical studies, reviews, proposals, and so on and so forth. He has averaged about ten publications a year for over thirty years!

What does the future hold for Dennett? Well, he is not even at normal retirement age yet, so he is far from the end of his productive life. As we said, a new book on free will is being prepared even as we write. Beyond that, who knows? Dennett has written deep, ground-breaking books on all the topics that he set himself over thirty years ago and has lived in a sea of cut-and-thrust for his whole career as a result. (His public confrontations with John Searle and Jerry Fodor and especially Stephen Jay Gould are the stuff of legends.) But over and over he has shown that he still has new things to say about the topics that matter to him.

This ends the chronological summary of Dennett’s corpus. We turn now to the interesting project of fitting the pieces together into a coherent whole. We will start with how the man fits into his time. It is not by accident that Dennett has been so influential. As well as being a brilliant polemicist, he started work at a pivotal point in the study of cognition, the beginning of the cognitive revolution. His own work at the time was an important part of that pivot.

2. Dennett’s Philosophy

There are a number of ways in which a scholar or scientist can be influential outside their home discipline. One is to have a single large idea. Another is to develop an important new technique. A third is to be so penetrating that just about anything one says on any topic is of interest. It might be supposed that Daniel Dennett’s influence beyond philosophy has been of the third kind. His main analytic tool is probably the ingenious rhetorical question, the revealing intuition pump. Dennett’s two best-known works outside of professional philosophy, *Consciousness Explained* (1991a) and *Darwin’s Dangerous Idea* (1995), seem consistent with this diagnosis. Nonetheless, there is a distinctively Dennettian point of view and its influence has been more important than Dennett’s work as a critical gadfly, brilliant and entertaining as the latter often is.

This point of view is clear already in *Content and Consciousness (C&C),* a blueprint for Dennett’s entire corpus. One of the decisive moves in the cognitive revolution was the rejection of empiricism as it existed at mid-century. Dennett was one of those who charted a course beyond this form of empiricism and we will take this issue as our starting point.

2.1 Empiricism in Dennett’s time

As we said, Dennett was a student of the iconoclastic Oxford philosopher of mind, Gilbert Ryle, and, before that, of the Grand Admiral of logic and epistemology, W.V.O. Quine. Ryle wrote during the heyday of ‘ordinary language philosophy’, the methodological movement inspired by J.L. Austin and given a tremendous boost by a certain (mis)reading of Wittgenstein’s *Philosophical Investigations.* Ordinary language philosophers supposed, roughly, that some philosophical problems appear to be insoluble because philosophers misinterpret the language used to state them. Take a notion dear to Dennett’s heart, belief. We speak of someone’s *believing* that snow is white, and of someone’s *believing* in the Loch Ness monster, and of their cat’s *belief* that it’s suppertime. It is easy to overinterpret these remarks as implying that ‘belief’ must denote a kind of state or object, as ‘gas’ does. We then go searching for properties that could gather all and only beliefs into a well-behaved set of such states or objects. Terrible difficulties at once arise. No one (philosophers excepted) ever entertains to themselves the phrase ‘snow is white, by gum’, yet most believe that proposition if they believe anything. It is hard to see, however, how a typical person could be said to believe in the reality of the Loch Ness monster unless they *had* sometime entertained the phrase ‘Loch Ness monster’. As for the cat, its inability to entertain *any* phrases leaves its status as a believer-of-specific-contents puzzling: can something have beliefs about suppertimes if, as is likely, it is incapable of believing that supper comes after lunch? Questions like these led ordinary language philosophers to think that something had gone deeply wrong. Typical English speakers seem to use words such as ‘belief’ without feeling compelled to suppose that cats secretly talk to themselves, or that tiny signs bearing the words ‘snow is white’ flash inside people’s brains from time to time, or other absurdities that their words *seem* to imply.

This example of dissolving a philosophical puzzle is far too crude to represent the work of any actual ordinary language philosopher but it will suffice as an illustration of the method. The crucial move in it is to take the way people express themselves in language as *data*, i.e., as a basis for philosophical reflection. However, data can be used in more than one way. In *C&C*, at least initially Dennett started from the same data as ordinary language philosophers but found that they led him to very different conclusions. Dennett never lost his respect for what he learned from his mentors about dissolving philosophical pseudo-puzzles. However, he found in ordinary uses of mentalistic language support, not for bland behaviourism, but for a kind of evolutionary cognitivist, as it is now called. (The kind of evolutionary cognitivist in question is the view that to understand behaviour, we must invoke a rich repertoire of mental states and processes, and that we can best identify the states and processes concerned by viewing the mental as having the shape it has due to natural selection.) Dennett’s turn against his philosophical upbringing here was momentous; it marks one of the clearest turning points that we know of in the movement away from simple-minded behaviourism to the current (near)-consensus cognitivism.

Ryle’s chief conclusion in *The Concept of Mind* is that philosophical hyper-literalness about the mentalistic idiom had, over several centuries, given rise to an absurd metaphysics of mind. Mentality is not a part of physical objects; physical objects do not hope for things or fear things or believe things. But minds are not parts of objects of any other kind, either. Mental objects no more hope or fear or believe than do physical ones. *People* hope or fear or believe – but, Ryle gives us reason to think, it is not clear that the word ‘person’ is being used to refer to an object here, certainly not to an object by itself. Ryle demonstrates that ascription of mentalistic predicates to objects is a ‘category mistake’, a logical error akin to ascribing a colour to an abstraction (‘What colour is the square root of 3?’) or contents to a vacuum (‘It’s chock-full of Nothingness’). Dennett, in *C&C*, summarizes the logical problem involved in applying mentalistic predicates to objects and processes by reference to the concept of *intentionality.* (The concept of intentionality was originally identified by medieval philosophers and reintroduced into philosophy by the nineteenth-century philosopher-psychologist Brentano.) The mental is *about* other states of affairs, whereas physical objects or states of physical objects are not. (Thus, even a good candidate case for mindless about-ness, such as 'That tree's rings are about its age', is not one; it means, 'That tree's rings *show* its age'.) It is a category mistake to take anything having aboutness to be straightforwardly a property of an object of any kind.

It will be objected, "We *do* speak about the animate bodies as engaging in activities that are about something; we speak of creatures' activities as directed to achieving goals and goals are about whatever they seek. For example, we say that she is walking *to get to* the store or he is dancing *because* it is fun. And so on. If Ryle is right, how is this possible?" It is possible because we see the motions involved in cases like these as *actions*, that is to say, as expressing reasons for doing things. Reasons for action are not causes, not causes as we usually think of them in the Humean tradition at any rate. Now, how does this distinction between reasons and (usual) causes accounts for our ascribing goals and other things having intentionality to bodies (i.e., actions) without introducing metaphysical extravagances (mentalistic 'ghosts in the machine', to use Ryle's phrase)? Answer: only if we can see how our reasons for doing the actions we do and the undeniably present causes of the bodily motions in which those actions consist relate to one another. In ordinary language, they peacefully coexist. We know that if we wish our arm to go up, then, under normal circumstances, it will go up, but we need not have any explanation of how either the reasons for this action or the causes of the related motion work.

This peaceful coexistence breaks down when we start doing scientific psychology. If all events must have physical causes, then, since beliefs, etc., are not normal physical causes, the psychological theorist cannot invoke things like beliefs to explain behaviour. At this point, theorists are tempted to go in one of two directions. Either one may *identify* beliefs with states of the body (usually the brain), or one may deny that there are any such things as beliefs. The first path requires abandoning the everyday concept of belief and is otiose, since now the notion of belief is not doing any explanatory work that notions of causes in brain circuitry can’t do. Radical behaviourists who denied that minds exist spoke as if they favoured the second option, but they were just adding redundant metaphysical noises to their scientific talk, according to Ryle. That is, radical behaviourists seemed to be saying ‘We’re not content that it’s methodologically unnecessary for us to talk about these things; we insist that they don’t exist. We don’t need this claim; we just ...’ – what? Hate it when people use alternative language about things like beliefs? It is not among a scientist’s responsibilities to be a dogmatic metaphysician. (The *worst* move is to imagine that beliefs *cause* brain-states, for this multiplies entities gratuitously while *still* abusing the everyday concept.)

There is another alternative: logical behaviourism (the term usually applied to Ryle’s position). On this view, talk about the causes of behaviour and talk about the reasons for behaviour can co-exist because talk about reasons is a way of finding a particular *pattern in* behaviour, not a way of finding the *causes of* behaviour. On this view, scientific psychology and the mentalistic conceptual scheme are taken to have no direct connection to one another; it is only the philosopher’s unwarranted assumption that they *must* serve the same functions that forces the postulation of ‘the ghost in the machine’: an object-like mind that both furnishes reasons for actions *and* is the cause of behaviour.

This happy ecumenicism ceases to be any comfort as soon as someone adds the opinion that science is in the business of describing the world *as it actually is*. In that case, what matters is not whether mentalistic talk is logically compatible with talk about causes of behaviour; all that matters is that science finds no place for such talk. Dennett’s other teacher, Quine explicitly said what Ryle only implied, namely, that mentalistic language describes no facts and is merely a ‘dramatic idiom’. Quine marshalled an important point of logical analysis in support of this view. Scientific language, it is generally supposed, ascribes properties to objects (e.g., ‘Spacetime is curved’). If so, one ought to be able to interchange co-referring terms within scientific sentences without changing their truth-values. Thus, if ‘The morning star orbits the sun’ is true, then ‘The evening star orbits the sun’ must also be true, since ‘the morning star’ and ‘the evening star’ refer to the same object. However, this logical property, called '*substitutability salva veritate',* does not apply inside many sentences in mentalistic discourse. 'Dweezil believes that the morning star orbits the sun' is perfectly consistent with the falsehood of 'Dweezil believes that the evening star orbits the sun'. In uttering such clauses, one is committed only to someone's intentional state being about something believed in, desired, hoped for, dreaded or whatnot, not something that exists. By contrast, if one ascribes a property to chairs, one commits oneself to the existence of chairs.

So 'the unicorn in which John believes' refers (if it refers to anything) to what some philosophers call an 'intentional object', not to an actual animal. Unless, it seems, brain-states can be directly taken to somehow be about intentional objects, sentences describing mental states won't have truth-values that derive in any straightforward way from those of non-mentalistic sentences. For Quine, there is a ready explanation for this semantic peculiarity: sentences describing mental states do not literally *describe* anything. They are merely what Quine calls a dramatic idiom for something else. In both Quine's eyes and those of his friend Skinner, this analysis buttressed the scepticism of the radical behaviourist about the existence of the mental.

This was the high point of twentieth century empiricism. Nothing exists, it was thought, except what we can sense using the sensitive surfaces of the body (eyes, ears, nose, taste, touch) and the correlations of these inputs with behavioural outputs. There may in addition be some apparatus connecting inputs to outputs or there may not but it does not matter – no such apparatus would be mental, as the mental has always been understood. Such empiricism pervaded the social sciences and some of the natural ones as well. Anthropologists 'discovered' that humans consist of arbitrary, infinitely variable belief-systems incommensurable with one another. For sociologists, the pressures of class membership caused people to invent value-systems that are merely rationalizations of power. Economists turned their venerable axioms of rationality into summaries of observed choice-behaviour lacking any deeper justification. Political scientists focussed on patterns in voting and other political behaviour. Many biologists emphasized the extent to which species are chains of descent shaped by geological and other environmental accidents. In the case of the biologists, this might have been more spin than substance but the social sciences were dominated by the idea that contingent environmental forces drive all behavioural processes. This was the environment into which Dennett came as a student.

The above description suggests a degree of doctrinal uniformity and simplicity that of course never really existed; but notice that now, thirty-five years later, the picture has no aptness at all. Articles and monographs now pour forth in which social scientists urge their colleagues to concentrate on the universal features of human behavioural patterns shaped by natural selection and fixed in the structures of cognition. Among biologists, the remaining defenders of pervasive contingency as the motor of evolution snarl defensively at 'ultra-Darwinian' adaptationists who depict Mother Nature as an engineer carving a rationalizable trajectory through design space. Barkow, Cosmides and Tooby (1992), in their manifesto for evolutionary psychology, aim explicitly at the old empiricism, which they call 'the standard social science model' (SSSM) (the phrase is increasingly used as a rallying banner across the disciplines). The foundation of the empiricist SSSM was its concept of mind. In the cognitive revolution against it, Chomsky is the Copernicus. Jerry Fodor and Dan Dennett perhaps fight for the role of Galileo.

Unlike Chomsky and Fodor, Dennett did not overthrow the empiricist concept of mind wholesale. Ryle's analysis is not an inaccurate review of what people *do* in fact say about minds, and Quine's logical semantics remains the single most influential body of work in the field. Dennett showed that we can preserve Ryle's and some of Quine's leading insights while pushing our study of behaviour away from the periphery and deep into the organism.

2.2 The foundations of Dennett's system

So how did Dennett respond to the radical empiricism of his time? To answer this question, we need to learn more about how he thinks about the mind and explanations of behaviour in terms of mental states. In many of his works, he starts by distinguishing three explanatory stance that one can take toward a complex organism or system. Consider a simple chess-playing computer (a favourite example of Dennett's).

One stance is to explain its current behaviour and predict its future behaviour by understanding how it is built and what the causal processes in it are like. This would give us an extraordinarily detailed, secure explanation of the system but at a severe price – extreme complexity. Not even the programmers who wrote the chess-playing programme could give this kind of explanation of the system. This stance Dennett calls *the physical stance.*

Another stance is to predict and explain the system's behaviour by understanding the design built into it, in this case the programme controlling its operations. Dennett calls this *the design stance.* This stance will produce an explanation much simpler than the first one – all we need to understand is how the system is designed to behave, not all the details of how and how well it implements this design – but the simplicity comes at a price. For our explanations to be any good, the system has to function as it was designed to function – and we have to assume that it is doing so or the design stance is useless to us. Because of this assumption built into them, explanations from the design stance are less secure than explanations from the physical stance.

A third stance is to predict and explain the system's behaviour by treating it as having goals and some capacity to achieve these goals. We can ask, what is the system trying to do, and what means is it adopting in order to do so. For example, we see it move pawns from in front of its queen. We could adopt the design stance and ask, what rules did the programmers build into the programme to cause it to do this. Or we can ask, 'What is it trying to do? Ah, get its queen out early' and reason, 'It knows that it's got to open a path through its pawns in order to do so.' This Dennett calls *the intentional stance.*

The intentional stance has the advantage of great simplicity. Instead of worrying about thousands of causal relationships (physical stance) or hundreds of rules, qualifiers, etc., built into procedures (design stance), all we have to worry about it one little goal, one little desire, to get its queen out early, one little belief, that to do it has to open a path through its pawns, and some fairly simple supplementary beliefs about how to do so in a way that is consistent with the rules of chess and the position of other pieces on the board. A three or four line inference and we are there.

Of course, this simplicity comes at a cost: we have to treat the system, not just as having goals and something like beliefs about how to realize its goals (some will already find this too high a cost), but even more importantly, as having sufficient rationality to figure out, given what it believes, what it has to do to satisfy its goals. And note: these things have to be assumed or the intentional stance does not work, so we cannot use the intentional stance to test these attributions. Thus, the intentional stance buys its simplicity at the cost of accepting assumptions that it can treat a system not just as having mental contents but also a degree of (minimal, self-interested) rationality.[1]

We now introduce a point that is crucial to understanding Dennett's thought. The intentional stance has often been taken to be the centre of his point of view. To read him this way is to miss a rich forest due to paying too much attention to a couple of the trees in it. The intentional stance is just that – a stance. The intentional stance is merely part of an explanatory schema. The really original and interesting part of Dennett's work is found in his view of what the world that underlies that schema is like. What underlies the intentional stance and justifies it as an explanatory schema is a powerful, comprehensive view about the nature of intentionality and the role of natural selection (evolution) in its very coming to be. Dennett's response to the radical empiricism of his mentors, indeed his distinctive contributions as a whole, are to be found in this comprehensive view.

[1] Breaking explanation down into a series of nested kinds of explanation, ordered by the size of the basic unit of explanation (beliefs and desires, designed functions, physical components) became a major organizing strategy throughout cognitive studies generally in the 70s. The version that came to dominate cognitive science goes like this. First, describe how a system is behaving, what tasks it is performing (often done using the intentional stance). Then ask: What procedures, etc., would allow a system to do these things (an application of the design stance). Finally ask: How are these procedures, etc., implemented in the brain or the brain and its environment? (This is not exactly the physical stance, which would study the brain by itself, not how material of other stances is related to it, but related to the physical stance).

In *C&C*, Dennett sets up a ‘philosopher’s dilemma’, which we might with hindsight call ‘the empiricist’s dilemma’. The problem is this. On the one hand, Ryle had shown that the mentalistic conceptual scheme is not translatable into the language of mechanisms and Humean causes. Behaviourism thus seems to hold out the only hope for a coherent science of mind, as Ryle had implied and Quine had explicitly said. On the other hand, as Chomsky had already argued in his famous review of Skinner’s *Verbal Behaviour* in 1959, behaviourists themselves seem to be stuck with the language of intentionality. To take one of Dennett’s examples, a behaviourist has not left the ghost in the machine behind if she talks of a rat’s *seeking* food, but how can she even begin to understand the rat’s behaviour if she does not take it to have goals, a way of grouping stimuli as similar, and so on? (Judgments of similarity require a judgment that ‘this a is an F’, which is to say, a network of beliefs.) Can Dennett have his cake and eat it, too – can he preserve the intuition that we should work with what we can observe, namely, behaviour, and yet make room for our rich and, it would seem, vital talk about minds?

The key point, and one that Dennett saw as early as Chapter 4 of *C&C*, is that intentional states are ascribed as a result of an interpretation of relations among behaviour, environment and dispositions, under the constraint that the target system is sufficiently rational, given its beliefs, to act in ways appropriate to reaching its goals. This means that intentional states are not ascribed to parts of a person, even parts as large as the brain. Intentional states are states of the *whole person,* indeed a person/ environment whole. If intentional states are complex triangulations of behaviour/brain/environment interactions, then they cannot be reduced to brain states. So what status do properties of persons and environments as a whole attributed under the constraint of rationality have?

Recently, Dennett (2000) has admitted that his ontological intuitions about intentional states are in “happy disarray.” At one stage, he accepted the label of ‘instrumentalist’, the view that treats anything unobservable as merely a postulation useful for making predictions of future states given the current states that one is observing. On this view, unobservable states like beliefs and desires have as much reality as things like centres of gravity – and for instrumentalists, that is not very much. (Dennett [1993] recalls this episode of self-labelling with chagrin.) More recently, though not abandoning his Quinean zeal to reject superfluous entities such as belief and desire circuits in the brain, Dennett has disavowed the label. His certainty that there is *something* real about the patterns we spot using the intentional stance along with strong realist intuitions about the things we talk about when we use in particular the physical stance convinced him that 'instrumentalist' was an inaccurate label for his position. The issue has drawn him deeply into debates over the status of 'folk psychology' that have preoccupied philosophers since the late 1970s (1991b). (A term due to Dennett himself, folk psychology comprises our vast storehouse of beliefs, intuitions, maxims, etc., about how motives, beliefs, and perceptions of various kinds will issue forth in other motives and beliefs and in behaviour.) By 1987, the regularities described by folk psychology had become, for Dennett, a 'virtual machine' implemented in the relationships among the distributed neural hardware of the brain, the environment, and behaviour. (His speculative account of the role played by this machine in giving rise to the human 'self' in *CE* has become a crucial aspect of his developed view of agency, as we will see.) Whatever the ontological status of the mental, Dennett views folk psychology and the network of intentional concepts in which its generalizations are couched as irreducible and an indispensable piece of apparatus in the behavioural sciences, probably the majority opinion among philosophers of mind (Clark 1989). Who would have expected, thirty-five years on, that the most stubborn opposition to the 'eliminativists', as those who look forward to the disappearance (at least from science) of the intentional framework are called, would be coming from the foremost student of Ryle's?

Dennett's attitude to the empiricism in which he was steeped in his philosophical youth is to be found in this tricky business of nailing down the exact ontological status of mental states and events such as beliefs and desires (the 'ontological status' of something refers to whether it exists and if so, what sort of existence it has).

2.3 Free-floating intentionality and Darwinian reasons

Dennett's view that intentionality is *attributed* to systems in order to rationalize their behaviour has generated a large body of critical debate centred mainly around two consequences of the claim. First, it suggests that a particular agent's intentional states, including her beliefs and purposes, are not 'up to her', or even entirely determined by facts internal to her. Second, it raises a question whether ascriptions of intentionality are anything other than arbitrary projections of the ascriber's interpretation. For example, how do we decide that saying a rock has the 'goal' of holding paper in place would be entirely gratuitous, while saying that a bee has the goal of finding honey is not? The intuition seems clear, but we should like to know why. It is not because the rock is incapable of formulating its goal to itself, since the same is true of the bee searching for nectar. Uniquely among philosophers at the time he wrote *C&C*, Dennett finds a solution to both aspects of the problem in Darwin:

> The principles of evolution proposed to explain learning and discrimination in the brain have the capacity to produce structures that have not only a cause but also a reason for being. That is, we can say of a particular structure that the animal has it because it helps in certain specified ways to maintain the animal's existence. It is a structure for discriminating edible from inedible material, or for finding one's way out of danger. [C&C, p. 64]

Let us call the reasons thus produced Darwinian reasons. The mature Dennett might have amended one phrase here, changing "maintain the animal's existence" to 'maximize its expected number of offspring given its environment', since natural selection 'cares' about an animal's existence only derivatively, and only as long as it continues to have opportunities to replicate. Nonetheless, the passage is remarkable. Spelled out, the approach that Dennett recommends (in 1987b, for example) is to start by asking, What goals ought the organism to have, given its evolutionary history and the possibilities presented by its environment? And what beliefs ought it to have, given its evolutionary history, experience and cognitive apparatus? We then ask what in this repertoire of goals and beliefs would make what it is currently doing rational in its current environment?

Before we consider this method in its natural (i.e., biological) setting, let us revisit the case of relatively simple artificial problem-solvers such as the chess-playing computer discussed above. Where such machines are concerned, goals are relatively transparent, since they provide the justification for the deliberate engineering of the devices. Consider Dennett's (1971) famous example of the thermostat. We may make no scientific error if we refer to the thermostat as *believing* that it's getting colder and *wanting* to warm the room up and therefore *acting* to do so. However, one can convey more information just as briskly by saying that the thermostat was designed to register the temperature of the room and switch on the furnace whenever its mercury drops below a certain predetermined level. There is no *point* – not just given our purposes as explainers, but also given the *facts* about the system – in ascribing intentional states to it, and doing so is therefore a scientific error.

As soon as we move to even a somewhat more complicated system such as a chess-playing computer, however, the decision to use the intentional stance is forced. The computer's goal is to win its games. However, there are many ways of trying to do this and all of them must be sensitive to unpredictable contingencies in the trajectory of play. The successful chess programme must therefore be designed so as to be able to formulate subsidiary goals in response to information coming from the world. As a result, its behaviour must be driven by complex interactions of internal informational states. This still does not *force* us to treat the system as intentional. However, as its repertoire encompasses diverse paths to the goal that have little in common except *being paths to that goal*, the regularities in its behaviour become visible only from the intentional stance. This is more than economy of description. Explaining why the computer moved a piece by recounting the history of impulses through its circuits, rather than by noting that it would otherwise have faced check-mate, would fail to support the right counterfactuals; that is, it is not true that the computer would not have moved the piece had its micro-history been different, but it *is* true that it would not have moved the piece had it not been threatened. This is why the intentional stance tracks ‘real patterns’ (Dennett 1991b) and is not merely the dramatic idiom of radical Quinean empiricism. (We will return to intentional attribution in section 2.6).

With the chess-playing computer, we can adopt the design stance or the intentional stance as we choose. Its intentional states are easy to identify: we know they will all be about, and only about, chess. However, the fact that the computer happens to have been designed by engineers with conscious goals in mind is in itself irrelevant to whether its behaviour in fact *does* conform to the intentional patterns associated with chess. How did certain patterns come to be the intentional patterns for chess, the ones that constitute rational strategies, etc.? Suppose that playing and winning chess matches were crucially relevant to some animal’s reproductive fitness. Then, in time, natural selection might well hit upon the information-processing capacities necessary and sufficient for skilful play. In this case, according to Dennett, it would simply be arbitrary to withhold attributions of intentional states about chess to the animal. That is how patterns get to be intentional ones. Note that the information-processing and problem-solving capacities of actual animals, including humans, are precisely the products of such design pressures. This does not imply any attribution of consciousness to natural selection; you don’t need consciousness to achieve intentionality. This conclusion leads Dennett to the doctrine of ‘free-floating intentionality,’ that is, to the view that intentional contents are not strictly determined by the subjects whose intentions are at issue. Neither the chess-playing computer nor the bee has any concept of its own intentional states, but they have them nevertheless. Rocks, since they are not selected for anything, have none.

This naturalistic grounding of intentionality in Darwinian processes points to a major factual mistake in radical empiricism, the last departure by Dennett from the empiricism of his youth that we will note. Often the production of Darwinian reasons requires complex processes within brains. Not always; intentionality is so free-floating that selection pressures *could* achieved it in systems without complex information-processing capacities. But often is enough; in those cases, any behaviourist form of empiricism has to be wrong.

None of this entails that a system *cannot* become aware of its own pre-programmed goals and, through complex feedback mechanisms, modify them. Clearly, some organisms, especially humans, can override biological goals, including fundamental ones: people sometimes do, for example, decide not to have children. This does not require that some cluster of connections encouraging baby-making be unwired, for there is no such dedicated cluster. Rather, when the disposition to want to reproduce is ascribed to a person, what we are really ascribing is a host of micro-dispositions that tend to a single outcome, namely, reproduction – dispositions to flirt, compete for the attention of prospective mates, find certain people more attractive than others, and so on, together with some relational properties yoking these dispositions to the environment. A conscious decision to use these micro-dispositions for purposes disassociated from reproduction need not (indeed, generally does not) override them. The micro-structures still serve the original Darwinian reasons but the environment is manipulated by conscious choice so that 'normal' processes such as sexual intercourse fail to serve the goals for which they were selected. More simply, while a person can decide not to have sex, they cannot just decide not to want sex; achieving the latter requires, for example, monks to undertake difficult, complex, and not always successful self-reengineering.

Dennett occupied a good portion of the 1970s and 80s skirmishing with opponents of AI, such as Searle (1984, and then over and over again) and Dreyfus (1979), who objected that whereas computers have merely derivative intentionality, humans and (some) animals have it intrinsically. However, this objection ignores the point of the appeal to Darwinian reasons. Biological systems do not – indeed, as Dennett had stressed in C&C (p. 57), could not – arrive in the world as directionless assemblies of cells and then arbitrarily formulate goals for themselves. Young animals are born with a number of generic goals (find reliably dry, warm, nutrient-abundant environments) and highly specific ones (stay close to mother), all engineered by natural selection in service of the only goal to which it can be directly sensitive, maximization of the probability of making more young animals. Hence Dennett's (1987a) prima facie shocking claim that Mother Nature is the only locus of 'intrinsic' intentionality; there are, as Dennett puts it, no other ‘unmeant meaners’.

In the 1990s, a major dispute erupted over just how far Darwinian reasons can take us, the so-called Darwin Wars (**Ross 2**). Gould (1997) and others grouped Dennett together with Dawkins and Maynard Smith as ‘ultra-Darwinians’. What is at stake may be no less than a coherent view of humanity’s place in nature. Here is what the argument is about: Dennett, like Dawkins, maintains that the way to do biological science (including psychology and linguistics and, we would wish to add, economics [**Ross 1**]) is to take as one’s default assumption that what exists has occurred for Darwinian reasons, and to establish further hypotheses for investigation by asking what the specific Darwinian reasons in question would likely have produced. This is the position referred to in biology and elsewhere as *adaptationism.* Gould and his allies maintain the opposite view, taking as their default assumption that what exists is substantially accidental, and that where one does find Darwinian reasons at work, one should not base further hypotheses on this discovery because the rationale is likely to fail outside of the strict circumstances in which it has been discovered to operate. Consider an example of this last point.

A baboon might luckily kill a human assailant by throwing a dropped spear at him. Since baboons often defend themselves by throwing things, we would say that the monkey intentionally defended itself with a weapon. But if the same baboon shot someone with a pistol, we would say that this was a bizarre accident, rather than generalizing from the precedent of the spear. Note that this judgment turns on what we know about *patterns* in baboon behaviour. A typical baboon simply has had no opportunity for acquiring information about how guns work; and in the absence of any belief that pulling the trigger will deter his enemy, we have no grounds for inferring from the monkey’s lucky shot that he desired to fire. However, it is not implausible that if enough baboons observed enough such accidents, they might stumble upon the regularity. At that point, we would indeed have intentional marksmonkeys.

The idea that patterns in intentional response are often brought about by copying and other forms of social pressure has been especially significant to Dennett, since if this is not the case then either a highly implausible sort of biological determinism must hold or intentionality is not really free-floating in the case of people. For this reason Dennett has made extensive use of Dawkins's (1982, 1989) concept of memes. Memes are the cultural analogues of genes, strings of informational code that pass easily from one mind to another, and are thus the basic units on which cultures are built. According to Dennett, they are also the basic units on which individual human personalities, or 'selves', are built; for selves, as essentially social creations, are expected to be individualized *assemblages* out of a stock of comprehensible pieces, that is, memes. (To see this, compare 'He has tastes I don't share: military music and pro wrestling' with 'He is crazy: he likes music by artists whose names begin with 'B' and sports that are played in the rain.' In the first case, you can understand the person's tastes, though they may seem obnoxious or whatever, but in the second case, you're just baffled. How could anyone have tastes like *those?*

Memes have frequently been the object of derision by philosophers and biologists who deplore careless analogies, since there are many properties (though the exact list is hotly disputed; see Dennett 1995) of genes that memes do not have. We are not sure how one can evaluate the aptness of an analogy except by its fruitfulness. Whether one does or does not approve of the *word* meme, Dennett's point, for which his entire corpus of writings constitutes a sustained argument, remains: minds are constituted by small semantic units, and the basic semantic units, call them what you like, cannot be deliberately chosen or willed into being. There is, of course, an important sense in which memes *are* human inventions. But what makes an idea a meme, something that may contribute to the self-narrating of many people, is its catchiness, not its content per se. And catchiness is a complex property built on relationships amongst contents, environments (composed most significantly of other memes, which is why intellectual history *is* tremendously important) and biological predispositions. It is thus not a property that is typically within the control of individuals. Here then is another, now social, respect in which intentionality is held by Dennett to be 'free-floating'.

The importance of Dennett's ruminations on these matters cannot be underestimated. They are what have carried his influence well beyond the usual territory of philosophers of mind and psychology. All sorts of things spin off from his approach to intentionality and we will spend the remainder of this introduction exploring some of them.

2.4 How do we come to have an intentional stance?

If natural selection can give us reasons for action, namely, Darwinian reasons, how do we come by *the ability to recognize and ascribe* these reasons? How do we come by folk psychology, aka the intentional stance? Maybe the turn to Darwin can help us here, too. The learning task involved in acquiring folk psychology is *prima facie* formidable. Folk-psychological characterizations, we have seen, are system-level rationalizations of processes to whose causal basis the part contributed by the brain is largely concealed. It seems highly unlikely, therefore, that the rich internal logic of folk psychology could have arisen spontaneously in the course of our trying to make sense of our world. Furthermore, if as many suggest folk psychology has not substantially changed in thousands of years – has not even *drifted*, like other large-scale cultural inventions such as poetry or warfare –, that too suggests that it is not a learned theory of behaviour. The obvious alternative is to suggest that, as many theorist hold to be true of language, it is carried innately. Dennett (1996) offers a speculative history of the evolution of folk psychology consistent with this hypothesis, and a micro-industry within developmental psychology and cognitive ethology is currently arguing for the existence of such as capacity in children (**Griffin and Baron-Cohen**) and testing the limits of such a capacity in adult primates (**Seyfarth and Cheney**). Wellman (1990) surveys the child developmental literature, Allen and Bekoff (1997) the animal literature.

The nature and extent of the innate component in folk psychology is still very much an open question. Certainly, some innate dispositions such as a wish to cultivate smiles and hugs from parents and to use language to communicate according to cooperative maxims are necessary for learning *any* social coordination behaviour, including the folk-psychological framework within which we understand such behaviour. However, as McGeer (1996) has argued, it may also be sufficient. Parents treat their children as apprentice members of society before they can talk and everyone around a child is continuously treating everyone else as rational agents. Poverty-of-stimulus arguments (arguments that children encounter very little and relatively degenerate evidence from which to construct a language) provide powerful support for the view that syntax is innate. It is not obvious that a similar argument is available for folk psychology. One alternative possibility is that where organisms with language and a few other innate social dispositions are concerned, adopting the intentional stance is a solution on which ordinary learning has to converge because there is no other. On the other hand, the best current theory of autism does provide some support for the view that folk psychology, or at least important aspects of its basis, may be innate. The best current theory of autism is that autistic people lack the implicit theory of intentional systems that most of us have, so their ability to adopt this stance to themselves and others is impaired or absent (**Griffin and Baron-Cohen**). Since this deficit seems to be inherited, the ability to adopt the intentional in those who have it may also be innate.

2.5 A theory of consciousness

If Dennett want intentionality to 'float free', he faces a challenge. That is not at all how matters *feel* to people. We experience ourselves as consciously directing much of our intentional life; much of the content of one's mental life is felt to be under one's direct and private control. Since Freud, many people accept that they have some beliefs and desires that behave strangely but they are called 'unconscious'. In the realm of consciousness, mental content does not seem to 'float'. It seems to be fixed in the private experience of the conscious self. Most research on consciousness sides with these common intuitions. Dennett cannot. If these intuitions are correct, then Dennett’s theory of intentionality is wrong and the ghost remains in the machine. It is thus incumbent on Dennett to conceptualize conscious phenomena in a way that squares with his view of intentionality yet also does justice to common experience. He rose to this challenge with enthusiasm in *Consciousness Explained* (1991a).

For Dennett, the appropriate method to study consciousness is just a special application of the general method of interpretation of behaviour in order to ascribe mental states given to us by the intentional stance. On the new application, we seek to interpret, not what mental states makes sense of behaviour, but how things will *seem* to someone given their behaviour. Dennett (1991a, pp. 72-9) calls this *heterophenomenology,* ‘hetero’ because it is done by another, ‘phenomenology’ because it seeks to construct how things seem to someone, i.e., what their phenomenology is like. When we pay attention to how things seem to ourselves, we are using the same method, interpreting how things must seem to someone given how they are behaving, but, because it is applied to self, not other, and because it uses whatever information is available from the inside, not just behaviour, Dennett calls it *autophenomenology.* However, it is just as much a matter of interpretation and just as much *not* a matter of ‘reading off’ independently-existing features as heterophenomenology.

Conscious states or what many philosophers call *qualia*, the felt quality of things, are thought-experiment heaven. ‘Thought experiment’ is another name for ‘intuition pump’ and mounting thought experiments is one of the most characteristic activities of the contemporary philosophers who study consciousness. We find stories about Mary the Colour Blind Scientist, the population of China as a single unified consciousness, inverted qualia, absent qualia (i.e., zombies – a system functioning like a mind while not conscious), and so on. Dennett fights fire with fire. He tells stories about CADblind computers (systems that can not just generate but also interpret and react to images, yet would still be unconscious on the story of mysterians philosophers such as Nagel and Searle), zimboes (zombies who don’t cheat and thus behave *exactly* like us, right down to rich responsiveness to others’ behaviour and capacity to monitor and report on their own states), and so on. One of the most famous is the Case of the Lady Remembered with Glasses:

> A short time after seeing a lady who was in fact not wearing glasses, we remember seeing her with glasses. When did the mistake occur? It seems that there should be two possibilities: the lady in question might have been misperceived with glasses or she might have been perceived correctly but, due to a later revision, immediately misremembered.

Dennett calls the first option a Stalinesque option because the way events are portrayed was a show from the beginning, the second an Orwellian option because the initial portrayal was accurate but history almost immediately gets rewritten. And his point is, if you make the time interval small enough, no possible information could decide for one rather than the other of these 'options' – in which case, they are not real options and never were. Put another way, when the error began is something we can stipulate but there is no fact of the matter about it (1991a, p. 125). If so, conscious states are not anything remotely like *qualia* as philosophers have thought of them – discrete, distinguishable 'felt qualities' that we can introspect with perfect accuracy (**Brook)**. Whether there is no such thing as felt qualities at all is another question (Brook 2000).

If the Cartesian picture of conscious states is wrong, what about the Cartesian theatre meant to house these states? Here is the basic problem with it:

> The solitary audience in the theatre of consciousness, the internal decision-maker and source of volitions or directives, the reasoner, if taken as *parts* of a person, serve only to postpone analysis. The banishment of these concepts from our analysis forces the banishment as well of a variety of other self-defeating props, such as the brain-writing waiting to be read, the mental images to be seen, the volitions to be ordered and the facts to be known. [*C&C*, 190].

If in addition conscious *states* are not neat, discrete little mental events, then what is the system that has them like? As we said in section 1., Dennett calls his substitute model the Multiple Draft

Model of consciousness. Its main features include:

1. content-fixing processes (particularly judgments). These processes do not generate clearly discrete content-bearing states, however.

2. multiple drafts. These are drafts, versions, of narrative-fragments, etc., that the content-fixing processes generate.

3. procedures for picking out certain drafts. These procedures select some drafts, deselect others, and tie some of the selected drafts together in a variety of usually loose, sometimes well-integrated ways. The results are temporary 'heads of mind' that Dennett calls *virtual captains.* The virtual captain of the present moment is me; and the story that comes to be told in me in which the succession of virtual captains hang together roughly coherently is the self (or, if you prefer, the Self).

4. probes. Among the procedures for selecting drafts and thereby giving them the cognitive prominence that is consciousness, auto-stimulation by probes into one's current information-bearing states is perhaps the most important.

5. memory storage. These contents, especially ones that take charge and become part of a 'head of mind' of any duration, get remembered; such memory is criterial for consciousness, Dennett insists (1991a, p. 132)

6. implementation. All this could be implemented in something like the old Pandemonium architecture proposed by Selfridge (1959).

In the workspace, packages can come to take charge of the system by commanding computational resources, in particular the system's resources for probing itself and linking what it finds to other information packages and to behaviour. Dennett urges that, unlike the Cartesian theatre, MDM allows for temporal issues in connection with consciousness to be blurry, a matter of interpretation and decision, just as they have turned out to be. Because the MDM sees consciousness as at bottom a matter of content, almost of programming, some theorists have serious reservations about it (**Churchland**).

When an alliance of information packages controls the self-monitoring and self-expressive capacities of the system, it becomes a short-term 'virtual captain' of the system. Appearances to the contrary, such a virtual captain is what we are talking about when we talk about the Self. Though Dennett has not been entirely clear on this, what links the current virtual captain to earlier virtual captains to produce the sense each of us has of continuing over time is memory, specifically the kind of autobiographical memory in which we remember earlier experiences from the point of view of having them, emotions from the point of view of feeling them, actions from the point of view of doing them. (If we expressed what we remembered, we would say things like, 'I remember seeing ...', 'I remember feeling ...', 'I remember doing ...')

Of the many interesting questions about MDM, here we will ask just one, because it sheds crucial light on Dennett's whole approach to consciousness. What is the status of MDM? Does Dennett see it as a description of what the various phenomena of consciousness consist in, or does he see MDM as a picture of a system that could contain consciousness? This is an important question. On the first alternative, though accepting that consciousness as such exists, Dennett would then be advocating that we eliminate many long-accepted features of consciousness from our theory as not existing – not only would there be no *qualia* of the sort that philosophers have believed in, there would be no unified, conscious subject of experience of the kind assumed by the tradition either. By contrast, on the second alternative all Dennett would be committed to is that the system that implements the various phenomena of consciousness is quite different from what we might have expected. He need not be saying anything about what consciousness itself is like.

This question has been put to Dennett. His response is very interesting:

> Am I am eliminativist? I am a deflationist. The idea is to chip the phenomena of the mind

> down to size, undoing the work of inflationists who actively desire to impress upon themselves and everybody else just how supercalifragilisticexpialidocious consciousness is, so that they can maintain, with a straight face, their favourite doctrine: The Mind is a Mystery Beyond All Understanding. [2000, p. 369-70]

Dennett is not merely telling an implementation story. Rather, *by* telling a neurally realistic, uninflated story about what the system that implements consciousness is probably like, he hopes to bring us to see more clearly what the phenomena being implemented must really be like.

It is central to Dennett's picture of consciousness that we can could learn about consciousness by inferences from behaviour and other externally accessible evidence. This idea is profoundly disturbing to a number of philosophers, notably Nagel (1974, 1991). Consciousness, according to these philosophers, is essentially subjective and ineffable, and therefore not the sort of thing on which any sort of third-person investigation could shed light. There is some irony in these quarrels between the subjective essentialists and Dennett. In effect, they turn his *conceptual* distinction between the domain of reasons and the domain of causes into a divide between our *epistemological access* to the subjective and the objective and reject all uses of behavioural evidence in studying the internal dynamics of consciousness. Dennett goes far in the opposite direction; not just behaviour but also neuroscience can illuminate consciousness.

There is a social dimension to Dennett's view of consciousness and the conscious self. The intentional stance is a coordinating device, essential for successful life with others. Those who fail to learn it are deemed autistic and children are incited almost from birth to use it to interpret others – and themselves. Applied to self, the intentional stance not only provides one with an account of one's reasons for doing this or that particular action, it provides one with a sense of oneself as a continuing being with a coherent history and unified opening to the future. This narrative sense of self allows for the distinctively human aspect of consciousness, as opposed to the mere discrimination of information and the meta-representation of certain brain-contents to assist in guiding behaviour, and it arises from one taking the intentional stance toward oneself. This self-creating story is also the source of our sense of ourselves as agents, according to Dennett; this idea has not, to put it mildly, achieved universal acceptance. Notice that Dennett's account of the genesis of selves is possible only if people do *not* have direct access to the informational states of their brains, and thus must predict and explain their own behaviour using the same resources as are available to third parties.

2.6 Uses of Dennett's methods

Dennettian method in behavioural science has three main parts. First, we need a rigorous and disciplined way of formulating models of agents as rational interactors. This move is very much like game-theoretic microeconomics (**Ross 1**). We then apply 'reverse engineering', using whatever knowledge we can gather about the constraints under which selection occurred (biological, cultural or both as the case may warrant) to frame hypotheses about differences between the behaviour of the 'ideally' rational agents identified in the first move and the actual products of real histories.

As well as leading to a highly original theory of consciousness, this methodology yields fruit in fields far removed from traditional philosophy. We have already mentioned one: autism research. As we said, the best current theory of autism is that autistic people lack the implicit theory of intentional systems that most of us have, so their ability to adopt this stance to themselves and others is impaired or absent (**Griffin and Baron-Cohen**).

Another is cognitive ethology. In section 2.2, we illustrated the application of the intentional stance using computers and other artefacts. In fact, reflection on our ways of understanding artificial systems is not the best way to explore the methodology of intentional stance reasoning. In computers, goals are fairly obvious, since devices are engineered precisely to achieve certain goals. The challenges facing the intentional stance are clearer in cognitive ethology, specifically the kind of cognitive ethology in which researchers use the intentional stance to construct models of animal minds on the basis of their behaviour. Natural observation alone will generally not suffice so researchers must mount experiments. Where subjects have language, experimental design is massively facilitated by the fact that subjects can be asked questions. If we want to know whether a subject is discriminating certain information along a particular dimension, we need merely ask her to perform a (typically verbal) task that requires that discrimination.

Where non-human animals and human infants are concerned, however, we cannot do that. In these cases, we must carefully work out a battery of probable Darwinian reasons at work; we could not probe animal minds by *any* means if we had no idea as to what sorts of things they pay attention to and why. Suppose, for example, that one is studying baboons. Since we know that their social structures are matrilineal, we may reasonably expect that males, at least before they are integrated into a troop, face the problem of getting females to pay attention to them. Now, suppose we are wondering whether the rare fights between males are motivated by intent to intimidate, drive away or harm the rival, or are bouts of showing-off designed to impress females. It would of course be suggestive if we passively observed that males fight only when females were looking, but merely suggestive. After all, females are almost always around in baboon society, and might be expected to have their attention drawn to loud squabbles regardless of whether these were being held for their benefit. Alternatively, perhaps male baboons do not get sufficiently excited, in general, in the absence of females to have their tempers riled to combat pitch. We need a Dennettian probe here, perhaps as follows. While males are fighting or displaying at each other, we might distract the females with something the males cannot see, thus encouraging the females to look away from the combatants but perhaps leading the latter to believe that the females have simply lost interest in their fracas. Do the males, at this point, go on fighting or break off and try to recruit attention? Of course, such an experiment could not be definitive, but no experiment, in any field, ever is. Do enough work of this sort, however, in sequences arranged to permit consilience, and one can gradually build a convincing picture of some aspects of the animal's *persona* (**Seyfarth and Cheney**; see also their 1990). (Note that we could not even interpret this imaginary experiment unless we have grounds for non-arbitrarily assigning beliefs and desires to the baboons.)

We have already mentioned that game-theoretic microeconomics makes moves very much like those in the first part of Dennett's method, though few economists are probably aware of any connection between their work and Dennett's. This parallel may have far-reaching implications (**Ross 1**).

And what about computers? We just said that computers are not the best way to illuminate the intentional stance, but what about the reverse? Is the intentional stance useful in artificial intelligence research? This is a complicated question that we will merely mention here. (**Wilks** gives a somewhat pessimistic assessment.)

2.7 Two general worries

There are at least two big, systemic worries that have been expressed about Dennett's work. One starts here. In (1988b), he provided a commentary on lessons for philosophers of mind derived from his experiences in Kenya with Seyfarth and Cheney. These may be summarized as follows. Much philosophy in the empiricist tradition has been forced by its starting assumptions to conclude that languageless animals cannot have minds (except metaphorically), since minds are taken to be constituted by propositional attitudes and propositions must be tokened by sentences that express their content. (For a sophisticated defence of this view, see Davidson 1975.) In *C&C*, Dennett rejects this conclusion on the following grounds.

The basis of the intentional stance is, as we have seen, the fact that internal cognitive states of some organisms and artefacts produce patterns in information tracking, manipulation and behaviour-control that can be identified only from the stance. But since particular ascribed intentional states need not, indeed generally do not, map onto discrete states of the brain or brain and behaviour, the absence of sentential tokens need not disqualify a creature from being a genuine bearer of intentional states. What could Seyfarth and Cheney have been investigating? The answer cannot be merely vervet *behaviour*, since they did not need their experiments to investigate behaviour; they could simply observe that. And they were certainly not studying vervet *brain-states*. They tell us that they are studying the *minds* of the monkeys, 'how they see the world', and that is the only remaining candidate.

On the other hand, in (1996) and elsewhere, Dennett has argued with great persuasiveness that language is central to our kind of mind. He has been known to speculate, presumably not entirely serious about the exact number, that a mind with a language like ours will be 20,000 times more intelligent than one without. For one thing, minds with language can learn from work that has gone before and been written up. For another, while minds without language are largely restricted to learning by trial-and-error, which can be lethal if you pick the wrong trial, minds with language can 'let their hypotheses die in their stead', to use Karl Popper's memorable phrase. So where does Dennett stand on the relationship of mindedness and language? The issue requires an essay all to itself to resolve (**Clark).**

A second general worry about Dennett's approach is nicely displayed in a dialogue between Churchland and Ramachandran (1993) and Dennett (1993). The worry is that the intentional stance and its application by way of Darwinian and other reasons for action seems to impose an *a priori* interpretation on the facts, rather than letting the facts 'speak for themselves'. The dialogue centres on the interpretation of so-called 'filling in' phenomenon in visual (and other) perception.

As you stare at a visual scene, you could be questioned (by others or by yourself) about billions of bits of information available to you – about relative depths, sizes, positions, colours, edges, shades, angles, grains, contours or complex combinations of these – and you could respond correctly more or less at once and without reflection. When you so respond, obviously you are not consulting a stored mental copy of the scene. For one thing, your brain lacks the requisite storage capacity, for the recursive-combinatorial power of our shared descriptive capacities pushes the informational carriage towards infinity ('the bush behind the tree six feet away from the rock parallel to the fence running past the creek'). Furthermore, your eyes are at any time receiving information from, and likely have *ever* received information from, only a fraction of the total scene on which you can report. You cannot report on a particular flower or bug in your visual field until someone or something directs your attention to it. On the other hand, you can close one eye without a hole popping into the field of the opposite eye's blind spot and the scene remains a stable array despite the continuously darting saccades of your eyes. In general, what you are literally transducing is a jumping, patchy, gap-filled sequence of partial scenes and when your brain presents this as a single, stable scene, it is evidently doing a great deal of 'correcting'. In what does this 'correcting' consist?

Much of the literature implicitly suggests that the brain constructs a coherent scene by producing mental models in which areas corresponding to unexamined regions of the external space are *filled in* by extrapolation from features of the examined regions. Thus, if you are looking at an expanse of plaid wallpaper, your brain might construct a representation of a uniformly plaid image by 'painting' the entire representation of the wall plaid based on the assumption that the surface of the wall is uniform. However, Dennett argues, this 'painting' metaphor is a residue of the Cartesian theatre, and must be unpacked in a way that does not involve reference to an inner presentation or observer. The brain need be troubled only by *actually observed* discontinuities. If it expects no information from unobserved regions, it need register *nothing* about them. If consistency is the system's default assumption, then when a subject is probed, or probes herself, she will *judge* that the wall is 'all plaid' without examining it further.

Churchland and Ramachandran argue that the experimental evidence is against Dennett here. They present a number of experiments that suggest that how things appear to subjects depends on the brain's having supplied non-transduced content in way that requires that non-transduced regions have been 'filled in' , rather than merely ignored. More generally, they complain,

> Part of the trouble with Dennett's approach to the various filling in phenomena is that he confidently prejudges what the neurobiological data at the cellular level will look like. Reasoning more like a computer engineer who knows a lot about the architectural details of the device in front of him than like a neurobiologist who realizes how much is still to be learned about the brain, Dennett jumps to conclusions about what the brain does not need to do, ought to do, and so forth ... While Dennett's idea ... may seem to have some engineering plausibility, it is really a bit of *a priori* neurophysiology gone wrong. Biological solutions, alas, are not easily predicted from reasonable engineering considerations. What might, from our limited vantage point, have the earmarks of sound engineering strategy, is, often as not, out of kilter with the way Nature does it. [Churchland and Ramachandran 1993, 42].

Dennett in his reply acknowledges that the evidence *might* force the conclusion that his conclusions are wrong. However, he does not budge with respect to the soundness of his method. It is, he says (Dennett 1993, 209), "a good way to think – risky, but a fine way of generating hypotheses [though] I might sometimes forget to admit, in the thrill of the chase, that these hunches need confirmation (and hence court disconfirmation)" (**Akins**).

This stand-off, more generally the relationship of Dennett's whole approach to neuroscience, is far from resolved. Moreover, it is related to the standoff between Dennett and Gould with respect to adaptationism that we examined earlier. We have argued that biological and social science could not get anywhere without *some* suppositions about Darwinian reasons. So perhaps all that Churchland, Ramachandran and Gould are saying is that Dennett lets the intentional stance do too much work. But how much is *too* much? Clearly we need inferences from the ‘logic’ of evolution and current behaviour-in-environment. Equally clearly, we need all the help we can get from actual palaeontological and other history, neuroscience, etc. Dennett has provided a powerful motive for theoreticians to get to work on sorting out the right balance. Indeed, he has done some of the needed work himself (1987c).

This concludes our introduction to the intentional stance and the philosophical system that Dennett has used it to generate. For the rest of the book, it will be taken as read.

3. To come

The essays to come explore the many areas beyond traditional philosophy in which Dennett’s work has had an important influence. **Brook** and **Churchland** explore Dennett’s contributions to consciousness studies, which now reach well beyond academic philosophy to include experimental psychology, AI, and neuroscience. (The first group of essays being about consciousness does not violate Dennett’s dictum that one should start with content, then take up consciousness because the introductory essay just completed give a full account of Dennett’s picture of mental content.) The second group of essays explores applications of the intentional stance in three very different domains. **Griffin and Baron-Cohen** explore Dennett’s contributions to developmental psychology and specifically our understanding of autism, **Seyfarth and Cheney** to cognitive ethology, and **Ross** to how one might understand economic theory. Next **Clark** and **Akins** explore the two interesting worries about Dennett’s work that were just introduced, the one about the relationship between language and his theory of content (also his theory of consciousness) **(Clark)** and the one about the rationalizing *a priorism* of the intentional stance and the facts as revealed by contemporary neuroscience (**Akins**). Finally, **Wilks** explores Dennett’s role in artificial intelligence research and **Ross** takes up his contributions to evolutionary theory and examines some of the deeper issues underlying the Darwin Wars of the 1990s.

The papers in this volume fall into two categories. Where the nature of a given influence is fairly straightforward but the influencing theory is complicated or possibly problematic, authors focus on the influencing theory. This is true of Brook, Churchland, Akins, and Clark. Where, on the other hand, the details of how Dennett has influenced a field are interesting in their own right, authors focus on the nature of the influence. This is true of Griffin and Baron-Cohen, Seyfarth and Cheney, Wilks, and Ross's second paper on the Darwin Wars. Ross's first paper has a foot in both camps.

References

Allen, C. and M. Bekoff (1997). *Species of Mind*. Cambridge, MA: MIT Press /A Bradford Book.

Barkow, J. L., Cosmides and J. Tooby, eds. (1992). *The Adapted Mind*. New York: Oxford University Press.

Brook, A. (2000). Judgements and Drafts Eight Years Later. In Ross, et al. 2000.

Chomsky, N. (1959). Review of Skinner, *Verbal Behaviour*. *Language* 35:26-58.

Churchland, P. S. and V. Ramachandran (1993). Filling In: Why Dennett is Wrong. In B. Dahlbom, ed., *Dennett and His Critics*. Oxford: Blackwell.

Clark, A. (1989). *Microcognition*. Cambridge, MA: MIT Press /A Bradford Book.

Davidson, D. (1975). Thought and Talk. In S. Guttenplan, ed., *Mind and Language*. Oxford: Oxford University Press.

Dawkins, R. (1982). *The Extended Phenotype*. Oxford: Oxford University Press.

Dawkins, R. (1989). *The Selfish Gene* (2nd edition). Oxford: Oxford University Press.

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

Dennett, D. (1971). Intentional Systems. *The Journal of Philosophy* 68:87-106.

Dennett, D. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting.* Cambridge, MA: MIT Press/A Bradford Book.

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

Dennett, D. (1987b). True Believers. In Dennett, 1987a.

Dennett, D. (1987c). Three Kinds of Intentional Psychology. In Dennett, 1987a.

Dennett, D., ed. (1987d). *The Philosopher's Lexicon, 8th edition.* American Philosophical Association Publication.

Dennett, D. (1988a). Quining qualia. *Consciousness in Contemporary Society*, A. Marcel and E. Bisiach, eds. Oxford: Oxford University Press

Dennett, D. (1988b). Out of the Armchair and into the Field. *Poetics Today* 9:205-221.

Dennett, D. (1991a). *Consciousness Explained*. Boston: Little Brown.

Dennett, D. (1991b). Real Patterns. *The Journal of Philosophy* 88:27-51.

Dennett, D. (1993). Back From the Drawing Board. In B. Dahlbom, ed., *Dennett and His Critics*. Oxford: Blackwell.

Dennett, D. (1995). *Darwin's Dangerous Idea*. New York: Simon and Schuster.

Dennett, D. (1996). *Kinds of Minds*. New York: Basic Books.

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

Dennett, D. (2000). With a Little Help From My Friends. In Ross, et al. 2000.

Dennett, D. and D. Hofstadter, eds. *The Mind's I.* New York: Basic Books.

Dreyfus, H. (1979). *What Computers Can't Do* (2nd edition). New York: Harper and Row.

Gould, S. J. (1997). Darwinian Fundamentalism. *New York Review of Books* 12/6/97.

McGeer, V. (1996). Is "Self-Knowledge" an Empirical Problem? Renegotiating the Space of Philosophical Explanation. *Journal of Philosophy* 93(10): 483-515.

Nagel, T. (1974). What is it Like to be a Bat? *Philosophical Review* 83:435-450.

Nagel, T. (1991). What We Have in Mind When We Say We're Thinking. *Wall Street Journal* 11/7/91.

Ross, D., A. Brook and D. Thompson, eds. (2000). *Dennett's Philosophy: A Comprehensive Assessment*. Cambridge, MA: MIT Press /A Bradford Book.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Searle, J. (1984). *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

Selfridge, O. (1959). Pandemonium: A Paradigm for Learning. *Symposium on the Mechanization of Thought Processes*. London: HM Stationary Office.

Seyfarth, R., and D. Cheney (1990). *How Monkeys See the World*. Chicago: University of Chicago Press.

Wellman, H. (1990). *The Child's Theory of Mind*. Cambridge, MA: MIT Press /A Bradford Book.