Oxford Handbooks Online

Intentional Systems Theory

Daniel Dennett

The Oxford Handbook of Philosophy of Mind
Edited by Ansgar Beckermann, Brian P. McLaughlin, and Sven Walter

Print Publication Date: Jan 2009 Subject: Philosophy, Philosophy of Mind
Online Publication Date: Sep 2009 DOI: 10.1093/oxfordhb/9780199262618.003.0020

Abstract and Keywords

Intentional systems theory is in the first place an analysis of the meanings of such everyday ‘mentalistic’ terms as ‘believe’, ‘desire’, ‘expect’, ‘decide’, and ‘intend’: the terms of ‘folk psychology’ that we use to interpret, explain, and predict the behaviour of other human beings, animals, some artefacts such as robots and computers, and indeed ourselves. In traditional parlance we seem to be attributing *minds* to the things we thus interpret, and this raises a host of questions about the conditions under which a thing can be truly said to have a mind, or to have beliefs, desires, and other ‘mental’ states. According to intentional systems theory, these questions can best be answered by analysing the logical presuppositions and methods of our attribution practices, when we *adopt the intentional stance* toward something.

Keywords: intentional systems theory, folk psychology, mentalistic terms, human behaviour, mental states, intentional stance

INTENTIONAL systems theory is in the first place an analysis of the meanings of such everyday ‘mentalistic’ terms as ‘believe’, ‘desire’, ‘expect’, ‘decide’, and ‘intend’: the terms of ‘folk psychology’ (Dennett 1971) that we use to interpret, explain, and predict the behaviour of other human beings, animals, some artefacts such as robots and computers, and indeed ourselves. In traditional parlance we seem to be attributing *minds* to the things we thus interpret, and this raises a host of questions about the conditions under which a thing can be truly said to have a mind, or to have beliefs, desires, and other ‘mental’ states. According to intentional systems theory, these questions can best be answered by analysing the logical presuppositions and methods of our attribution practices, when we *adopt the intentional stance* toward something. Anything that is usefully and voluminously predictable from the intentional stance is, by definition, an *intentional system*. The intentional stance is the strategy of interpreting the behaviour of an entity (person, animal, artefact, whatever) by treating it *as if* it were a rational agent who governed its ‘choice’ of ‘action’ by a ‘consideration’ of its ‘beliefs’ and ‘desires’. The scare-quotes around all these terms draw attention to the fact that some of their standard connotations may be set aside in the interests of exploiting their central features: their role in practical reasoning, and hence in the prediction of the behaviour of practical reasoners.

19.1 The Three Stances

The distinctive features of the intentional stance can best be seen by contrasting it with two more basic stances or strategies of prediction: the *physical stance* and the (p. 340) *design stance*. The physical stance is simply the standard laborious method of the physical sciences, in which we use whatever we know about the laws of physics and the physical constitution of the things in question to devise our prediction. When I predict that a stone released from my hand will fall to the ground, I am using the physical stance. In general, for things that are neither alive nor artefacts the physical stance is the only available strategy, though there are important exceptions, as we shall see. Every physical thing, whether designed or alive or not, is subject to the laws of physics and hence behaves in ways that in principle can be explained and predicted from the physical stance. Whether the thing I release from my hand is an alarm clock or a goldfish, I make the same prediction about its downward trajectory, on the same basis. Predicting the more interesting behaviours of alarm clocks and goldfish from the physical stance is seldom practical.

Alarm clocks, being designed objects (unlike the stone), are also amenable to a fancier style of prediction: prediction from the design stance. Suppose I categorize a novel object as an alarm clock. I can quickly reason that *if* I depress a few buttons just so, *then* some hours later the alarm clock will make a loud noise. I don't need to work out the specific physical laws that explain this marvellous regularity; I simply *assume* that it has a particular design—the design we call an alarm clock—and that it will function properly, as

designed. Design-stance predictions are riskier than physical-stance predictions, because of the extra assumptions I have to take on board: that an entity *is* designed as I suppose it to be, and that it will operate according to that design; that is, it will not malfunction. Designed things are occasionally misdesigned, and sometimes they break. (Nothing that happens to, or in, a stone counts as its malfunctioning, since it has no function in the first place, and if it breaks in two, the result is two stones, not a single broken stone.) When a designed thing is fairly complicated (a chainsaw in contrast to an axe, for instance) the moderate price one pays in riskiness is more than compensated for by the tremendous ease of prediction. Nobody would prefer to fall back on the fundamental laws of physics to predict the behaviour of a chainsaw when there was a handy diagram of its moving parts available to consult instead.

An even riskier and swifter stance is the intentional stance, a subspecies of the design stance, in which the designed thing is treated as an agent of sorts, with beliefs and desires and enough rationality to do what it ought to do given those beliefs and desires. An alarm clock is so simple that this fanciful anthropomorphism is, strictly speaking, unnecessary for our understanding of why it does what it does, but adoption of the intentional stance is more useful—indeed, well-nigh obligatory—when the artefact in question is much more complicated than an alarm clock. Consider chess-playing computers, which all succumb neatly to the same simple strategy of interpretation: just think of them as rational agents who *want* to win, and who *know* the rules and principles of chess and the positions of the pieces on the board. Instantly your problem of predicting and interpreting their behaviour is made vastly easier than it would be if you tried to use the physical or the design stance. At any moment in the chess game, simply look at the chessboard and draw up a list of all the legal moves available to the computer **(p. 341)** when its turn to play comes up (there will usually be several dozen candidates). Now rank the legal moves from best (wisest, most rational) to worst (stupidest, most self-defeating), and make your prediction: the computer will make the best move. You may well not be sure what the best move is (the computer may ‘appreciate’ the situation better than you do!), but you can almost always eliminate all but four or five candidate moves, which still gives you tremendous predictive leverage. You *could* improve on this leverage and predict in advance exactly which move the computer will make—at a tremendous cost of time and effort—by falling back to the design stance and considering the millions of lines of computer code that you can calculate will be streaming through the CPU of the computer after you make your move, and this would be much, much easier than falling all the way back to the physical stance and calculating the flow of electrons that results from pressing the computer's keys. But in many situations, especially when the best move for the computer to make is so obvious it counts as a ‘forced’ move, you can predict its move with well-nigh perfect accuracy without all the effort of either the design stance or the physical stance.

It is obvious that the intentional stance works effectively when the goal is predicting a chess-playing computer, since *its* designed purpose is to ‘reason’ about the best move to make in the highly rationalistic setting of chess. If a computer program is running an oil refinery, it is almost equally obvious that its various moves will be made in response to its

detection of conditions that more or less dictate what it should do, given its larger designed purposes. Here the presumption of *excellence or rationality of design* stands out vividly, since an incompetent programmer's effort might yield a program that seldom did what the experts said it ought to do in the circumstances. When information systems (or control systems) are well designed, the rationales for their actions will be readily discernible, and highly predictive—whether or not the engineers that wrote the programs attached ‘comments’ to the source code explaining these rationales to onlookers, as good practice dictates. We needn't know anything about computer programming to predict the behaviour of the system; what we need to know about is the rational demands of running an oil refinery.

19.2 The Broad Domain of the Intentional Stance

The central epistemological claim of intentional-systems theory is that when we treat *each other* as intentional systems, using attributions of beliefs and desires to govern our interactions and generate our anticipations, we are similarly finessing our ignorance of the details of the processes going on in each other's skulls (and in our own!) and relying, unconsciously, on the fact that to a remarkably good first approximation people are rational. We risk our lives without a moment's hesitation when we (p. 342) go out on the highway, confident that the oncoming cars are controlled by people who want to go on living and know how to stay alive under most circumstances. Suddenly thrust into a novel human scenario, we can usually make sense of it effortlessly, indeed involuntarily, thanks to our innate ability to see what people ought to believe (the truth about what's put before them) and ought to desire (what's good for them). So second-nature are these presumptions that when we encounter a person who is blind, deaf, self-destructive, or insane we find ourselves unable to adjust our expectations without considerable attention and practice.

There is no controversy about the fecundity of our folk-psychological anticipations, but much disagreement over how to explain this bounty. Do we learn dozens or hundreds or thousands of ‘laws of nature’ along the lines of ‘If a person is awake, with eyes open and facing a bus, he will tend to believe there is a bus in front of him’ and ‘Whenever people believe they can win favour at low cost to themselves, they will tend to cooperate with others, even strangers’, or are all these rough-cast laws generated on demand by an implicit sense that these are the rational responses under the circumstances? In favour of the latter hypothesis is the fact that whereas there are indeed plenty of stereotypic behaviour patterns that can be encapsulated by such generalizations (which might, in principle, be learned *seriatim* as we go through life), it is actually hard to generate a science-fictional scenario so novel, so unlike all other human predicaments, that people are simply unable to imagine how people might behave under those circumstances. ‘What would *you* do if that happened to you?’ is the natural question to ask, and along with such unhelpful responses as ‘I'd probably faint dead away’ comes the tellingly normative ‘Well, I hope I'd be clever enough to see that I should … ’. And when we see characters behaving oh so cleverly in these remarkably non-stereotypical settings, we have no difficulty understanding what they are doing and why. Like our capacity to understand entirely novel sentences of our natural languages, our ability to make sense of the vast array of human interactions bespeaks a generative capacity that is to some degree innate in normal people.

We just as naturally and unthinkingly extend the intentional stance to animals, a non-optional tactic if we are trying to catch a wily beast, and a useful tactic if we are trying to organize our understanding of the behaviours of simpler animals, and even plants. Like the lowly thermostat, as simple an artefact as can sustain a rudimentary intentional-

stance interpretation, the clam has its behaviours, and they are rational, given its limited outlook on the world. We are not surprised to learn that trees that are able to sense the slow encroachment of green-reflecting rivals shift resources into growing taller faster, because that's the smart thing for a plant to do under those circumstances. Where on the downward slope to insensate thinghood does ‘real’ believing and desiring stop and mere ‘as if’ believing and desiring take over? According to intentional-systems theory, this demand for a bright line is ill-motivated.

(p. 343) 19.3 Original Intentionality versus Derived or ‘as if’ Intentionality

Uses of the intentional stance to explain the behaviour of computers and other complex artefacts are not just common; they are universal and practically ineliminable. So it is commonly accepted, even by the critics of intentional-systems theory, that such uses are legitimate, so long as two provisos are noted: the attributions made are of *derived* intentionality, not *original* or *intrinsic* intentionality, and (hence) the attributions are, to one degree or another, *metaphorical*, not literal. But intentional-systems theory challenges these distinctions, claiming that

> **(1)** there is no principled (theoretically motivated) way to distinguish ‘original’ intentionality from ‘derived’ intentionality, and
> **(2)** there is a continuum of cases of legitimate attributions, with no theoretically motivated threshold distinguishing the ‘literal’ from the ‘metaphorical’, or merely ‘as if’, cases.

The contrast between original and derived intentionality is unproblematic when we look at the paradigm cases from everyday life, but when we attempt to promote this mundane distinction into a metaphysical divide that should apply to all imaginable artefacts, we create serious illusions. Whereas our simpler artefacts, such as painted signs and written shopping lists, can indeed be seen to derive their meanings from their functional roles in *our* practices, and hence not have any *intrinsic* meaning independent of *our* meaning, we have begun making sophisticated artefacts such as robots, whose trajectories can unfold without any direct dependence on us, their creators, and whose discriminations give their internal states a sort of meaning *to them* that may be unknown to us and not in our service. The robot poker player that bluffs its makers seems to be guided by internal states that function just as a human poker player's intentions do, and if that is not original intentionality, it is hard to say why not. Moreover, our ‘original’ intentionality, if it is not a miraculous or God-given property, must have evolved over the aeons from ancestors with simpler cognitive equipment, and there is no plausible candidate for an origin of original intentionality that doesn't run afoul of a problem with the second distinction, between literal and metaphorical attributions.

The intentional stance works (when it does) whether or not the attributed goals are genuine or natural or ‘really appreciated’ by the so-called agent, and this tolerance is crucial to understanding how genuine goal-seeking could be established in the first place. Does the macromolecule *really* want to replicate itself? The intentional stance explains what is going on, regardless of how we answer that question. Consider a simple organism —say a planarian or an amoeba—moving non-randomly across the bottom of a laboratory dish, always heading to the nutrient-rich end of the (p. 344) dish, or away from the toxic end. This organism is seeking the good, or shunning the bad—its own good and bad, not those of some human artefact-user. Seeking one's own good is a fundamental feature of any rational agent, but are these simple organisms seeking or just ‘seeking’? We don't

need to answer that question. The organism is a predictable intentional system in either case.

By exploiting this deep similarity between the simplest—one might as well say most mindless—intentional systems and the most complex (ourselves), the intentional stance also provides a relatively neutral perspective from which to investigate the differences between our minds and simpler minds. For instance, it has permitted the design of a host of experiments shedding light on whether other species, or young children, are capable of adopting the intentional stance—and hence are higher-order intentional systems. A *first-order* intentional system is one whose behaviour is predictable by attributing (simple) beliefs and desires to it. A *second-order* intentional system is predictable only if it is attributed beliefs about beliefs, or beliefs about desires, or desires about beliefs, and so forth. A being that can be seen to act on the *expectation* that you will *discover* that it *wants* you to *think* that it doesn't *want* the contested food would be a fifth-order intentional system. Although imaginative hypotheses about 'theory of mind modules' (Leslie 1991) and other internal mechanisms (e.g. Baron-Cohen 1995) have been advanced to account for these competences, the evidence for the higher-order competences themselves must be adduced and analysed independently of these proposals about internal mechanisms, and this has been done by cognitive ethologists (Dennett 1983; Byrne and Whiten 1988) and developmental psychologists, among others, using the intentional stance to design the experiments that generate the attributions that in turn generate testable predictions of behaviour.

The intentional stance is thus a theory-neutral way of capturing the cognitive competences of different organisms (or other agents) without committing the investigator to overspecific hypotheses about the internal structures that underlie the competences. (A good review of the intentional stance in cognitive science can be found in three essays in Brook and Ross (2002)—by Griffin and Baron-Cohen, Seyfarth and Cheney, and Ross.) Just as we can rank-order chess-playing computers and evaluate their tactical strengths and weaknesses independently of any consideration of their computational architecture, so we can compare children with adults, or members of different species, on various cognitive sophistications in advance of having any detailed hypotheses about how specific brain differences account for them. We can also take advantage of the intentional stance to explore models that break down large, sophisticated agents into organizations of simpler subsystems that are themselves intentional systems, subpersonal agents that are composed of teams of still simpler, 'stupider' agents, until we reach a level where the agents are 'so stupid that they can be replaced by a machine'—a level at which the residual competence can be accounted for directly at the design stance. This tactic, often called 'homuncular functionalism', has been widely exploited in cognitive science, but it is sometimes misunderstood. See Bennett and Hacker (2003) for objections to this use of the intentional stance and Dennett (2007) for a rebuttal. (p. 345) Hornsby (2000) offers a more nuanced discussion of the tensions between the personal and subpersonal levels of explanation.

A natural reaction to the intentional stance's remarkable tolerance of penumbral or metaphorical (or, to some critics, downright false) attributions is to insist on hunting for an essence of belief (and desire, etc.) that some of these dubious cases simply lack. The task then becomes drawing the line, marking the necessary and sufficient conditions for true believers. The psychologist David Premack (1983), for instance, has proposed that only second-order intentional systems, capable of beliefs *about* beliefs (their own and others') can really be counted as believers, a theme that bears similarity to Davidson's claims (e.g. 1975) about why animals are not really capable of *thought*. A more elaborately defended version is Robert Brandom's attempt to distinguish ‘simple intentional systems’ (such as all animals and all existing artefacts, as well as subpersonal agencies or subsystems) from ‘interpreting intentional systems’ in *Making it Explicit* (1994). Brandom argues that only social creatures, capable of enforcing norms, are capable of genuine belief. An alternative is to turn the issue inside out, and recognize that the ‘true’ cases are better viewed as limiting cases, extreme versions, of an underlying common pattern. Consider a few examples of the use of intentional terms, spread across the spectrum:

A. When evolution *discovers* a regularity or constancy in the environment, it designs adaptations that tacitly *presuppose* that regularity; when there is *expectable* variation instead of constancy, evolution has to go to the expense of specifying the adaptive response to the various different conditions.
B. When a cuckoo chick hatches, it *looks for* other eggs in the nest, and if it finds them, it *tries* to push them out of the nest because they are in competition for the resources it needs. The cuckoo doesn't understand this, of course, but this is the *rationale* of its behaviour.
C. The computer pauses during boot-up because it *thinks* it is communicating with another computer on a local area network, and it is *waiting for* a response to its greeting.
D. White castled, *in order to protect* the bishop from an *anticipated* attack from Black's knight.
E. He swerved because he *wanted* to avoid the detached hubcap that he *perceived* was rolling down the street towards his car.
F. She *wanted* to add baking soda, and *noticing* that the tin in her hand said ‘Baking Soda’ on it, she *decided* to open it.
G. Holmes recalled that whoever rang the bell was the murderer, and, observing that the man in the raincoat was the only man in the room tall enough to reach the bell rope, he deduced that the man in the raincoat was the culprit, and thereupon rushed to disarm him.

The last example is a paradigm of rational belief, but even in this case the attribution leaves a great deal of the reasoning inexplicit. Notice, too, that in the other cases of quite unproblematic, unmetaphorical human belief, such as (E), the swerving case, it is unlikely that anything like an *explicit representation* of the relevant beliefs and (p. 346) desires (the propositional attitudes) occurred in the driver's stream of consciousness. Had it not been for his beliefs about the relative danger of swerving and being hit by an object, he would not have taken the action he did; he would not have swerved to avoid a sheet of paper, for instance, and he would not have swerved had he believed there was a bus in the lane he swerved into, but it is not clear how, if at all, these guiding beliefs are *represented* unconsciously; there are so many of them. Similarly, the cook in (F) may have quite 'unthinkingly' opened the tin of baking soda, but she would not have opened a tin of loose tea or molasses had her hand fallen on it instead. Attributing a large list of *beliefs* to these agents—including propositions they might be hard-pressed to arteculate—in order to account for their actions is a practice as secure as it is familiar, and if some of these informational states don't pass somebody's test for genuine *belief*, so much the worse for the claim that such a test must be employed to distinguish literal from metaphorical attributions. The model of beliefs as sentences of Mentalese written in the belief box, as some would have it, is not obligatory, and may be an artefact of attending to extreme cases of human belief rather than a dictate of cognitive engineering.

19.4 Objections Considered

Although the earliest definition of the intentional stance (Dennett 1971) suggested to many that it was merely an instrumentalist strategy, not a theory of real or genuine belief, this common misapprehension has been extensively discussed and rebutted in subsequent accounts (Dennett 1987, 1991, 1996). The fact that the theory is maximally neutral about the internal structures that accomplish the rational competences it presupposes has led to attempted counter-examples.

(1) *The Martian marionette* (Peacocke 1983). Suppose we found an agent (called 'The Body' by Peacocke) that passed the intentional-stance test of agency with flying colours but proved, when surgically opened, to be filled with radio transceivers; its every move, however predictable and explicable by our attributions of beliefs and desires to *it*, was actually caused by some off-stage Martian computer program controlling the otherwise lifeless body as a sort of radio-controlled puppet. The controlling program 'has been given the vast but finite number of conditionals specifying what a typical human would do with given past history and current stimulation; so it can cause The Body to behave in any circumstances exactly as a human being would' (Peacocke 1983: 205).

This is no counter-example, as we can see by exploring the different ways the further details of the fantasy could be fleshed out. If the off-stage controller controls this body and no other, then we were certainly *right* to attribute the beliefs and desires to the person whose body we have surgically explored; this person, like Dennett in 'Where am I?' (Dennett 1978), simply keeps his (silicon) brain in a non-traditional location. If, on the other hand, the Martian program has more (p. 347) than one (pseudo-)agent under

control, and is coordinating their activities (and not just providing in one place *n* different independent agent-brains), then the Martian program *itself* is the best candidate for being the intentional system whose actions we are predicting and explaining. (The Martian program in this case really is a puppeteer, and we should recast all the *only apparently* independent beliefs and desires of the various agents as in reality the intended manifestations of the master agent. But of course we must check further on *this* hypothesis to see if the Martian program is in turn controlled by another outside agent or agents or is autonomous.) What matters in the identification of the agent to whom the beliefs and desires are properly attributed is autonomy, not specific structures. Of course a bowl of structureless jelly or confetti is not a possible seat of the soul, simply because the complex multi-track dispositions of a mind have to be realized somehow, in an information-processing system with many reliably moving (and designed) parts.

(2) *The giant look-up table* (Block 1982): Having used the intentional stance to attribute lots of clever thoughts, beliefs, and well-informed desires to whoever is answering my questions in the Turing test, I lift the veil and discover a computer system that, when opened, turns out to have nothing in it but a giant look-up table, with *all the possible* short intelligent conversations in it, in alphabetical order. The only 'moving part' is the alphabetic string-searcher that finds the next canned move in this pre-played game of conversation and thereupon issues it. Surely this is no true believer, even though it is voluminously predictable from the intentional stance, thereby meeting the conditions of the definition.

There are several ways of rebutting this counter-example, drawing attention to different foibles of philosophical method. One is to observe that the definition of an intentional system, like most sane definitions, has the tacit rider that the entity in question must be physically possible; this imagined system would be a computer memory larger than the visible universe, operating faster than the speed of light. If we are allowed to postulate miraculous (physics-defying) properties to things, it is no wonder we can generate counter-intuitive 'possibilities'. One might as well claim that when one opened up the would-be believer one found nothing therein but a cup of cold coffee balanced on a computer keyboard, which vibrated in just the miraculously coincidental ways that would be required for it to type out apparently intelligent answers to all the questions posed. Surely not a believer! But also not possible.

A more instructive response ignores the physical impossibility of the 'Vast' (Dennett 1995) set of alphabetized clever conversations, and notes that the canning of all these conversations in the memory is itself a process (an R&D process) that requires an explanation (unless it is yet another miracle or a 'cosmic coincidence'). How was the quality control imposed? What process exhaustively pruned away the stupid or nonsensical continuations before alphabetizing the results? Here it is useful to consider the use of the intentional stance in evolutionary biology:

> Francis Crick, one of the discoverers of the structure of DNA, once jokingly credited his colleague Leslie Orgel with ‘Orgel's Second Rule’: Evolution is cleverer than you are. Even the (p. 348) most experienced evolutionary biologists are often startled by the power of natural selection to ‘discover’ an ‘ingenious’ solution to a design problem posed by nature.
>
> When evolutionists like Crick marvel at the cleverness of the process of natural selection they are not acknowledging intelligent design! The designs found in nature are nothing short of brilliant, but the process of design that generates them is utterly lacking in intelligence of its own.
>
> (Dennett 2006: 37–8)

The process of natural selection is a blind, foresightless, purposeless process of trial and error, with the automatic retention of those slight improvements (relative to some challenge posed by the world) that happen by chance. We can contrast it with intelligent design. Now how did the giant look-up table consisting of all *and only* clever conversations come to be created? Was it the result of some multi-zillion-year process of natural selection (yet another impossibility) or was it handcrafted by some intelligence or intelligences? If the latter, then we can see that Block's counter-example is a close kin to Peacocke's. Suppose we discovered that Oscar Wilde lay awake nights thinking of deft retorts to likely remarks and committing these pairs to memory so that he could deliver them if and when the occasion arose ‘without missing a beat’. Would this cast any doubt on our categorization of him as an intelligent thinker? Why should it matter *when* the cogitation is done, if it is all designed to meet the needs of a time-pressured world in an efficient way? This lets us see that in the incompletely imagined case that Block provides it might *not* be a mistake to attribute beliefs and desires to this surpassingly strange entity! Just as Peacocke's puppet does its thinking in a strange *place*, this one does its thinking at a strange *time*! The intentional stance is maximally neutral about how (or where, or when) the hard work of cognition gets done, but guarantees that the work *is* done by testing for success. In the actual world, of course, the only way to deliver real-time cleverness in response to competitively variegated challenges (as in the Turing Test) is to *generate* it from a finite supply of already partially designed components. Sometimes the cleverest thing you can do is to quote something already beautifully designed by some earlier genius; sometimes it is better to construct something new, but of course you don't have to coin all the words, or invent all the moves, from scratch.

Coming from the opposite pole, Stich and others have criticized the intentional stance for relying on the rationality assumption, making people out to be much more rational than they actually are (see e.g. Stich 1981; Nichols and Stich 2003; Webb 1994). These objections overlook two facts. First, without a background of routine and voluminous fulfilment of rational expectations by even the most deranged human beings, such unfortunates could not be ascribed irrational beliefs in the first place. Human behaviour is simply not interpretable except as being in the (rational) service of some beliefs and desires or other. And second, when irrationality does loom large, it is far from clear that there is any stable interpretation of the relevant beliefs (see e.g. Dennett 1987: 83–116; 1994: 517–30).

(p. 349) 19.5 Summary

Intentional systems theory is a theory about how and why we are able to make sense of the behaviours of so many complicated things by considering them as agents. It is not directly a theory of the internal mechanisms that accomplish the roughly rational guidance thereby predicted. This very neutrality regarding the internal details permits intentional systems theory to play its role as a middle-level specifier of subsystem competencies (subpersonal agents, in effect) in advance of detailed knowledge of how they in turn are implemented. Eventually we arrive at intentional systems that are simple enough to describe without further help from the intentional stance. Bridging the chasm between personal-level folk psychology and the activities of neural circuits is a staggering task of imagination that benefits from this principled relaxation of the conditions that philosophers have tried to impose on (genuine, adult) human belief and desire. Intentional systems theory also permits us to chart the continuities between simpler animal minds and our own minds, and even the similarities with processes of natural selection that 'discover' all the design improvements that can thereby be discerned. The use of the intentional stance in both computer science and evolutionary biology, to say nothing of animal psychology, is ubiquitous and practically ineliminable, and intentional-systems theory explains why this is so.

References

Baron-Cohen, S. (1995), *Mindblindness: An Essay on Autism and Theory of Mind* (Cambridge, Mass.: MIT Press).

Bennett, M. R., and Hacker, P. (2003), *Philosophical Foundations of Neuroscience* (Oxford: Blackwell).

Block, N. (1982), 'Psychologism and Behaviorism', *Philosophical Review*, 90: 5–43.

Brandom, R. (1994), *Making it Explicit* (Cambridge, Mass.: Harvard University Press).

Brook, A., and Ross, D. (2002) (eds.), *Daniel Dennett* (Cambridge: Cambridge University Press).

Byrne, R., and Whiten, A. (1988) (eds.), *Machiavellian Intelligence: Social Expertise and the Evolution of Intelligence in Monkeys, Apes, and Humans* (Oxford: Oxford University Press).

Davidson, D. (1975), 'Thought and Talk', in S. Guttenplan (ed.), *Mind and Language: Wolfson College Lectures, 1974* (Oxford: Clarendon), 7–23.

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

—— (1978), *Brainstorms: Philosophical Essays on Mind and Psychology* (Cambridge, Mass.: MIT Press/Bradford).

—— (1983), ‘Intentional Systems in Cognitive Ethology: The “Panglossian Paradigm” defended’, *Behavioral and Brain Sciences*, 6: 343–90.

—— (1987), *The Intentional Stance* (Cambridge, Mass.: MIT Press).

(p. 350) Dennett, D. (1991), ‘Real Patterns’, *Journal of Philosophy*, 87: 27–51.

—— (1994), ‘Get Real’, *Philosophical Topics*, 22: 505–68.

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

—— (1996), *Kinds of Minds* (New York: Basic).

—— (2006), ‘The Hoax of Intelligent Design, and How It was Perpetrated’, in J. Brockman (ed.), *Intelligent Thought: Science Versus the Intelligent Design Movement* (New York: Vintage), 33–49.

—— (2007), ‘Philosophy as Naïve Anthropology’, in D. Robinson (ed.), Neuroscience and Philosophy: Brain, Mind, and Language (New York: Colombia University Press), 73–95.

Griffin, R., and Baron-Cohen, S. (2002), ‘The Intentional Stance: Developmental and Neurocognitive Perspectives’, in A. Brook and D. Ross (eds.), *Daniel Dennett* (Cambridge: Cambridge University Press), 83–116.

Hornsby, J. (2000), ‘Personal and Subpersonal: A Defence of Dennett's Early Distinction’, *Philosophical Explorations*, 3: 6–24.

Leslie, A. (1991), ‘The Theory of Mind Impairment in Autism: Evidence for a Modular Mechanism of Development?’ in A. Whiten (ed.), *Natural Theories of Mind* (Oxford: Blackwell), 63–78.

Nichols, S., and Stich, S. (2003), *Mindreading* (Oxford: Oxford University Press).

Peacocke, C. (1983), *Sense and Content* (Oxford: Oxford University Press).

Premack, D. (1983), ‘The Codes of Man and Beasts’, *Behavioral and Brain Sciences*, 6: 368.

Ross, D. (2002), ‘Dennettian Behavioural Explanations and the Roles of the Social Sciences’, in A. Brook and D. Ross (eds.), *Daniel Dennett* (Cambridge: Cambridge University Press), 140–83.

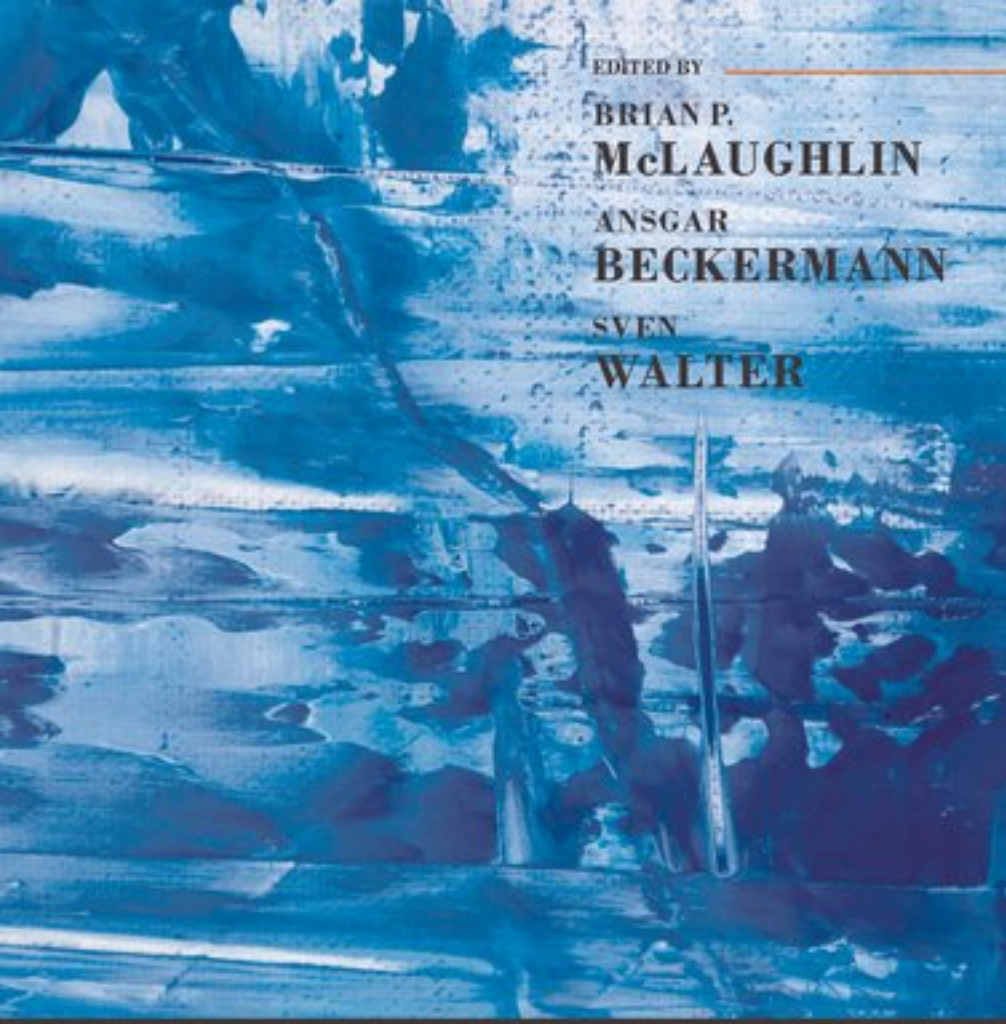
Seyfarth, R., and Cheney, D. (2002), ‘Dennett's Contribution to Research on the Animal Mind’, in A. Brook and D. Ross (eds.), *Daniel Dennett* (Cambridge: Cambridge University Press), 117–39.

Stich, S. (1981), ‘Dennett on Intentional Systems’, *Philosophical Topics*, 12: 39–62.

Webb, S. (1994), ‘Witnessed Behaviour and Dennett's Intentional Stance’, *Philosophical Topics*, 22: 457–70.

Daniel Dennett

Daniel C. Dennett is university professor and Austin B. Fletcher Professor of Philosophy at Tufts University. He is also the co-director of the Center for Cognitive Studies there. His most recent book on free will is Freedom Evolves (2003) and among his recent articles are “Toward a Science of Volition,” with W. Prinz and N. Sebanz, in Disorders of Volition, edited by N. Sebanz and W. Prinz (2006), and “Some Observations on the Psychology of Thinking about Free Will,” in Are We Free? Psychology and Free Will, edited by John Baer, James C. Kaufman, Roy F. Baumeister (OUP, 2008).



EDITED BY

BRIAN P. McLAUGHLIN

ANSGAR BECKERMANN

SVEN WALTER

The Oxford Handbook *of* PHILOSOPHY OF MIND