Daniel Dennett – True Believers: The Intentional Strategy & Why it Works

(Chapter 2 of The Intentional Stance)

Detailed Argument

Introduction

- Dennett supplies an introductory quotation from Somerset Maugham, entitled *Death Speaks*. A person has met Death in Baghdad market and interprets his gesture as threatening, so “escapes” to Samarra. It turns out that the gesture wasn’t a threat, but one of surprise, since Death had not expected to see the person in Baghdad, because he had an appointment with him that evening in Samarra¹.

- In the social sciences, there is lots of talk about *belief*, and talk about *talk about belief*. There is much controversy because belief is a curious, perplexing and multi-faceted phenomenon. Belief attribution is a complex business, especially for exotic, religious or superstitious beliefs. We court argument or scepticism when attributing beliefs to animals, infants, computers or robots. We’re uncomfortable attributing contradictory or wildly false beliefs to apparently healthy adult members of our own society. Could someone *really* believe that rabbits are birds? It would take quite a story to persuade us to attribute such a belief to someone.

- Attribution of problematic beliefs is beset by issues of subjectivity, cultural relativism and “the indeterminacy of radical translation”², whereas attribution in straightforward cases gives no trouble at all. When thinking of these straightforward cases, it almost seems possible in principle to confirm these simple, objective belief attributions by finding the beliefs themselves inside the believer’s head. You either believe X or you don’t believe X (taken to include no opinion), which is an objective fact about you that must come down to your brain being in a particular state. Hence, if we know enough physiological psychology, we could tell whether you believed there was milk in the fridge, whatever you said and however much you dissembled. On this view, physiological psychology could trump any “black box” method of the social sciences that divines beliefs by external criteria such as behavioural, cultural, social or historical.

- There are two extreme opposing views on the nature of belief and its attribution. While baldly stated, they are seen as mutually exclusive and exhaustive, so that the theorist can only be sympathetic to one of them. (1) *Realism* – having a belief is like being infected with a virus; a perfectly objective internal matter of fact which can often be reliably guessed at by an observer. (2) *Interpretationism*³ - the question of a person having a particular belief is analogous to whether the person is immoral or has style; it depends what you’re interested in, is relative and a matter of interpretation.

- Dennett thinks this dichotomy is a mistake. He is a realist in that he thinks beliefs are perfectly objective, but an interpretationist in that he thinks beliefs can only be discerned, and their existence confirmed, from the standpoint of a successful predictive strategy.

---

¹ Presumably, the point of this is the intentionality of the gesture.
² Of which Dennett appears scornful, noting that it requires phenomenological analysis, hermeneutics, empathy, verstehen, etc.
³ Dennett dislikes having to give this view a name.
• This intentional strategy – adopting the intentional stance – approximates to treating the object whose behaviour you want to predict as a rational agent with beliefs and desires exhibiting intentionality. Dennett will argue that any system whose behaviour is well predicted by this strategy is a believer in the fullest sense of the world. To be an intentional system is just what it is to be a true believer. Dennett has hitherto gained few converts, but will here deal with many compelling objections.

The Intentional Strategy and How it Works

• Dennett considers one of the “deplorably popular” methods of predicting behaviour – astrology – deplorable only because we have such good evidence that it doesn’t work, occasional success being due to luck or predictive vagueness. If it did work for some people, we could categorise them as astrological systems, whose behaviour was contingently predictable using the astrological strategy. If there were such people, we’d be interested in how the strategy worked – in its rules, principles and methods – and we could do that by asking astrologers, reading their books and watching them in action. But, we’d also be interested in why it worked, and we might find either no opinion on the part of the astrologers or pure hokum. So, having a good strategy and knowing why it works are two different things.

• Consider the physical stance, which allows behaviour prediction by determining a system’s physical constitution and the physical impingements on it and using the laws of physics, as in Laplace’s grand strategy for predicting the entire future of the universe. More modest versions work for predictions made by laboratory chemists and physicists, as well as cooks. While the strategy is not always practically available, that it will always work in principle, ignoring quantum indeterminacy, it is a dogma of the physical sciences.

• Sometimes it is more convenient to adopt the design stance, which ignores the messy physical details, and predicts that a system will behave as it is designed to do under the circumstances. Most computer users have no idea of the physical constitution of their machines, but can predict their behaviour, barring physical malfunction, based on what they are designed to do. Similarly one can predict the behaviour of alarm clocks based on a casual examination of their exteriors, or at a lower level, based on a description of the system of gears, without their material being specified.

• One can only predict the designed behaviour from the design stance – one would need to revert to the physical stance if one wanted to know what would happen if the clock were filled with liquid helium. Many biological objects – such as hearts or stamens – can be treated as designed as well as physical systems.

• Where even the design stance is practically inaccessible, one can retreat to the intentional stance. This works by (1) treating the object whose behaviour is to be predicted as a rational agent, (2) determining what beliefs the agent ought to have, given its purpose and place in the world, (3) determining by the same principles what desires it ought to have and finally, (4) predicting that the rational agent will act to further its goals in the light of its beliefs. In many cases, a little practical reasoning will give a decision about what the agent ought to do, and this is what we predict it will do.

• Dennett attempts to make the strategy clearer with a little elaboration. He asks how we populate one another’s heads with beliefs. He starts with truisms –
sheltered people tend to be ignorant, but exposure leads to knowledge. In general we come to believe all the truths we’re exposed to in our corner of the world. Sensory confrontation with x for a suitable period of time is normally sufficient for us to know or have true beliefs about x. We are highly suspicious of the claims to ignorance of those in a position to know (eg. that the gun was loaded).

- In fact, we only come to “know all about” and believe relevant truths, though anything interesting is learnable provided it is within my threshold of discrimination and the integration and holding power of my memory. Hence, one rule for attributing beliefs in the intentional strategy is to attribute all the beliefs relevant to the system’s interests and desires that its experience has made available. This has a couple of defects, in that we are forgetful even of important things and entertain false beliefs. However, Dennett thinks that the attribution of any false belief arises in the main from true beliefs, the falsehood starting from hallucination, illusion, misperception, memory loss, or even fraud – but that false beliefs grow in a culture of true beliefs.

- Dennett thinks that even arcane beliefs arise by a process of mainly good reasoning from beliefs already attributed. An implication of the intentional strategy is that true believers mainly believe truths; Dennett hazards that more than 90% of a person’s beliefs are arrived at using the rule in the bullet above.

- This rule is derived from a more fundamental one, namely to attribute to a system the beliefs it ought to have, and the same goes for desires. We attribute the usual list of basic beliefs to people (ie. survival, food, …) and citing such a desire terminates the “why ?” game of reason-giving. Trivially, we have a rule to attribute desires to a system for things it thinks good for it, and, less trivially, attribute to it desires for things it thinks the best means to other ends it desires. Attribution of bizarre or detrimental desires, like false beliefs, requires special stories.

- Verbal behaviour complicates the relation between beliefs and desires as the latter are attributed on its basis. We would find it difficult to attribute a desire for a meal specified in detail in the absence of verbal expression. Dennett thinks that language not only enables us to express complex desires but also to have them. Expressed desires are more particular than what would satisfy you. Once expressed, since you are a truth-teller, you are committed to their detailed satisfaction.

- One might object to being asked how many baked beans one wanted, but we are socialised to accede to similar requests we hardly notice and certainly don’t find oppressive. There is a parallel with beliefs, on which verbalisation forces an often unwanted precision. Focusing on the results rather than the effects of this social force can easily mislead us into thinking that obviously beliefs and desires are like

---

4 Dennett includes a long footnote discussing the differences amongst philosophers on this issue – ie. between those who think it obvious that most of our beliefs are true (Quine, Davidson, ...) and those who think it obviously false. His diagnosis is that they are talking about different things. Dennett suggests distinguishing between beliefs and opinions – the latter approximating to betting on the truth of a particular sentence. He considers Democritus’s beliefs – even though his physics was totally wrong, his ordinary beliefs (about where he lived, where to buy a good pair of sandals, …) were most likely true. Dennett counters the response that since all beliefs are theory-laden, a view he accepts and thinks important, and since Democritus’s theory was wrong, so were his quotidian beliefs. He asks why we should assume that Democritus’s explicit theory (his opinions) is what infects his daily beliefs rather than the same benign theory that undergirds the beliefs of his less sophisticated contemporaries? Democritus’s observational beliefs would be left largely untouched by change of his theoretical opinions, since few of them would be touched by it.
sentences stored in the head. Fully formed sentences that come true, or which we want to come true, may be unreliable models for the whole domain of belief and desire.

- With respect to the *rationality* of intentional systems, we start off charitably and revise downwards as circumstances dictate. We assume people are perfectly rational and believe all the implications of their beliefs and entertain no contradictions. The clutter of infinitely many implications isn’t a problem because we’re only interested in ensuring the system is rational enough to get to the implications relevant to its behavioural predicament of the moment. Dennett leaves aside, for this chapter, questions of irrationality and finitude, which he says raise particularly knotty problems of interpretation.

- Turning from description to use, Dennett claims people *use* the intentional strategy all the time, since it’s the only practical one we currently have, and that it *works* almost all the time. Dennett’s example is why we don’t think it a good idea for Oxford Colleges to dream up and award their own degrees whenever they feel like it. He doesn’t spell out the example but alludes to the sort of “what if” strategy we’d adopt in working out the probable consequences by thinking of how people would act. Because the intentional strategy is so habitual and effortless, we overlook the way it shapes our expectations of people – and of other mammals, birds, fish, reptiles and even shell-fish, for that matter. We devise traps to catch lesser creatures by reasoning about what they know, desire, avoid and so on. A chess-playing computer will not take your knight if it can see a reply that will let you win its rook. A modest thermostat will turn off the heater when it believes the room has reached the desired temperature.

- Some plants are *cautious* about *concluding* spring, which is when they *want* to blossom, has come early. Lightening always *wants* to find the quickest way to earth, but sometimes a clever electrician can *fool* it into *choosing* another path.

**True Believers as Intentional Systems**

- Dennett comes clean and admits that the quality of the belief attributions in the previous section varies from the serious to the dubious to the pedagogically useful metaphors to outright fraud. The next task maybe ought to be to distinguish those systems that *really* have beliefs from those that only appear to do so, but Dennett thinks this would be either a labour of Sisyphus or would just be terminated by fiat. The important thing to note is that even where we know the strategy works *for the wrong reasons*, at least it still does work, at least to a degree, and it is this that distinguishes the class of intentional systems, ie. from the class for which the strategy never works. However, is the latter class empty? Does the Oxford lectern from which Dennett is lecturing believe, like some of his auditors, that it is at the centre of the civilised world? Does it want to stay there, and adopt the best strategy of staying put. Is it therefore an intentional system, given that I can attribute beliefs and desires to it and predict its course of action? If so anything is.

- Dennett thinks that the lectern is disqualified because we already knew it was going to do nothing, and tailored its beliefs and desires in an unprincipled manner to suit. This isn’t the case with people, animals and computers, where the intentional strategy is the only strategy that works for predicting behaviour. It

---

5 This, of course, cries out for a sceptical response (but wait for the next section)!

6 I.e. never getting anywhere.
might be objected that this doesn’t reflect a difference in the nature of the systems, but only reflects our incompetence as scientists. Had we Laplacean omniscience, we’d be able to predict the behaviour of a computer or a human body (assuming it to be governed by the laws of physics) without recourse to the sloppy design and intentional strategies. Engineers manage to avoid anthropomorphising thermostats; their failure with more complex systems and artefacts is just symptomatic of human epistemic frailty, and we wouldn’t want to count them with ourselves as true believers on such parochial grounds. Wouldn’t it be intolerable for a system to be classified as a believer by one observer but not by a cleverer one? This would be radical interpretationism, which Dennett doesn’t accept, his view being that, while we are free to adopt the intentional stance or not, if we do the results of its adoption are perfectly objective.

- The success of the stance tends to be obscured by a focus on cases where it yields dubious or unreliable results. In chess, the intentional strategy, even when it fails to pick out just the move to be made, drastically reduces the possibilities from the full list of legal but bad moves.

- While we can’t predict the buy and sell decisions by stockbrokers or the exact speech to be given by a politician, we can successfully predict some of the sorts of decisions they will make or themes they will raise today. This lack of precision can be useful in allowing us to chain predictions. If the Secretary of State were to admit to being a communist agent, even though this would be such a startling event, we could still make many successful predictions, including chains of predictions. While mostly not startling, these predictions describe an arc of causation in space-time that could not be predicted by any imaginable practical extension of physics or biology.

- The intentional strategy’s power is illustrated by an objection of Nozick’s. Imagine some ultra-intelligent Martians to whom we are as thermostats are to clever engineers, so that they do not need the design or intentional stances to predict our behaviour as Laplacean super-physicists. While we might see brokers and bids on Wall Street, they see sub-atomic particles milling about and are such good physicists that they can predict days in advance the ink marks on the tape announcing the DJIA close. They can predict the behaviour of moving bodies without the need to treat them as intentional systems. Would we be right to conclude that from their perspective we aren’t intentional systems any more than thermostats? If so, our status as believers is not objective but is in the eye of ab beholder sharing our intellectual limitations.

- Dennett’s response is that if the Martians didn’t see us as intentional systems, then, despite their Laplacean predictions, they would be missing the perfectly objective patterns of human behaviour that support generalisations and predictions and are only describable from the intentional stance. If they see a stockbroker deciding to place an order, and predict the exact movements of phone-dialling finger and vibrating vocal-cords as he places the order, yet do not see that indefinitely many motions, even from different individuals, would have had the same impact on the market, then they have missed a real pattern in the world. One hasn’t understood how internal combustion engines or stock-markets work unless one realises the intersubstitutivity of one spark-plug for another, or one similar order for another. There are societal pivot points where what matters concerning what people do is whether they believe that p or desire A, and other similarities or differences between individuals are irrelevant.
• Dennett imagines a prediction contest between the Martian and an Earthling predicting a person’s actions based on a telephone call to the wife (I will turn up for dinner with the boss in an hour armed with a bottle of wine). Both predict the arrival of the car, but the Earthling’s prediction – a reasonable guess – seems like a miracle to the Martian, given the amount of calculation required (from the physical stance) and the Earthling’s obvious inability to perform them. Dennett claims that the coming true of the Earthling’s predictions would appear to someone without the intentionalist strategy as marvellous and inexplicable as the fatalistic inevitability of the rendezvous in Samarra. Dennett explains that fatalists (like astrologers) wrongly believe that the patterns in human affairs are inevitable and will transpire come what may, however much the victims scheme, second-guess and wriggle in their chains. They are almost right, in that there are patterns in human affairs – those we categorise in terms of the beliefs, desires and intentions of rational agents – which, while not quite inexorable, are capable of absorbing apparently random physical perturbations and variations.

• There is a cavil against this story, in that if the Martian is willing to enter into a contest with the Earthling, he must recognise the Earthling as an intentional system, and so, why not recognise all Earthlings as such, and the mystery would evaporate. Dennett imagines patching up the tale by stories of Earthling disguise, but thinks this would obscure the moral; namely, the unavoidability of the intentional stance with respect to oneself and one’s fellow intelligent beings. He admits that this is interest relative, in that one can adopt the physical stance to intelligent beings, oneself included. However, one must also adopt the intentional stance to oneself, and also to one’s fellows if one intends to learn what they know. Our Martians may fail to recognise us as intentional systems, but they must possess the relevant concepts because they view themselves as intentional systems if they observe, theorise, predict and communicate. The patterns are there to be described, whether or not we care to see them.

• Dennett stresses two things about the intentional patterns discernible in the activities of intelligent creatures; (a) their reality, the objective fact that the intentional strategy works as well as it does. (b) Since no-one is perfectly rational, unforgetful, observant or immune to fatigue, malfunction or design imperfection the intentional stance does not work perfectly, since these problems lead to situations it cannot describe. This is similar to the failure of the design stance to work with broken or malfunctioning artefacts. Even in the mild psychopathological cases of self-deception or the holding of contradictory beliefs, the intentional strategy fails to provide clear and stable verdicts on what desires and beliefs to attribute to such an one.

• A holder of a strong realist position on beliefs and desires would insist that the person in the above degenerate situation does have particular beliefs and desires, but that these are undiscoverable by the intentional strategy. Dennett, while not adopting a relativist, but rather a milder realist, position, thinks there is no fact of the matter in such cases, but that why and when this is the case is objective. He also allows the interest relativity of belief attribution, and that one culture might attribute different beliefs to a particular individual than would another. We can have multiple objective patterns with their respective imperfections and with

---

7 Cross-reference this in due course.
8 Dennett thinks that beings lacking these modes of action would be “marvellous, nifty and invulnerable” entities, but that we’d not call them intelligent. Quite so, which is why people don’t call computers intelligent.
objective facts about how well the respective intentional strategies work in predicting behaviour and out-gun rivals.

- The bogey of radically different and equally warranted interpretations derived using the intentionalist strategy is metaphysically important, though not if we restrict our attention to human beings, the most complex intentional systems we know.

- It’s not time to point out the obvious differences between ourselves and thermostats. What Dennett describes as the “perverse claim” is that all there is to being a true believer is being a system whose behaviour can be successfully predicted by the intentional strategy. In other words, that all there is to really and truly believing that p is being an intentional system for which p occurs as a belief in the most predictive interpretation. However, Dennett thinks that, in the context of interesting and versatile intentional systems, such an apparently shallow and instrumentalist criterion for belief puts severe constraints on the internal constitution of a true believer and consequently yields a robust version of belief.

- Considering the thermostat, at most we might attribute to it six beliefs – such as the room is too hot/cold, the boiler on/off, obtaining a warmer room requires it to turn the boiler on – and fewer desires. Suppose, given that it has no concept of heat and so on, that we de-interpret the thermostat’s beliefs and desires – believing that the A is too F, and so on, since by giving it different inputs and outputs it could regulate things other than temperature. As Dennett says, attachment to a heat-sensitive transducer and a boiler is too impoverishing a link to the world for us to grant any rich semantics to its belief-like states.

- Say we enhance its sensory inputs with eyes and ears with which to see and hear shivering and complaining occupants, and give it rudimentary geography to know the likely temperature on being told where it is. Dennett imagines us giving it other functions to perform, such as purchasing fuel, and generally enriching its internal complexity and giving its belief-like states more to do by providing more and different occasions for their deduction from other states and occasions to act as premises for further reasoning. The end result is to enrich the semantics of its dummy predicates so that it becomes less able to act as a maintainer of anything other than room temperature. That is, the class of indistinguishable satisfactory models of the formal system embedded in its internal states gets smaller and smaller as we increase complexity, until we get to a virtually unique semantic interpretation. Then, we would say that this device – or animal or person – has beliefs about heat and about this room because we cannot imagine any other context in which it would work.

- Our thermostat had beliefs about a particular boiler because it was fastened to it, but it could easily be attached to something else so that its minimal causal link to the world and the rather impoverished meaning of its internal states changes. For more perceptually rich and behaviourally more versatile systems, it becomes more difficult to substitute its links to the world without changing its organisation – it will notice the change in its environment and its internal states will change to compensate. Dennett claims that complex systems with fixed states require very specific environments to operate properly, but that those with states that are not fixed will adapt to the environment perceived by their sensitive sensory...

---

9 In justification, Dennett points out an analogy with cryptography – the more information available, the less likely there are to be radically different interpretations consistent with the data.

10 But, our reason for doubting it has beliefs is that it hasn’t anything to believe with.
attachments and be driven into a new operative state. The organism *mirrors* the environment, which is *represented* in the organisation of the system.

- Dennett stresses his views about the direction of belief attribution. When we find something for which the intentional strategy works, we try to interpret some of its internal states as internal representations. This is in contrast to attributing beliefs and desires only to things in which we find internal representations. What makes an internal feature of a thing into a representation has to be its role in regulating the behaviour of the intentional system.

- The reason Dennett has stressed our relation to the thermostat is his view that there is no magic moment as we move upwards from the thermostat to a system that *really* has an internal representation of its environment. He imagines a gradual transition from fancier thermostats to us via robots, each having a more demanding representation of the world than its predecessor. We are so intricately connected to the world that almost no substitution is possible except in thought experiment. Putnam imagines Twin Earth where everything appears to be an exact replica of earth, except below the threshold of our powers of discrimination – it is, in this case, what is called water on Twin Earth having a different chemical analysis to that on earth. Were you swapped with your Twin Earth replica you’d never be the wiser, just as was the simple thermostat with a much grosser change of inputs. For us, if Earth and Twin Earth aren’t virtual replicas, we will notice and change our states dramatically.

- Our beliefs are about our own boilers (rather than those on Twin Earth). Fixing the semantic referents of our beliefs requires facts about our actual embedding in the world. Dennett claims that the problems we have attributing belief to people are just the same as those we have attributing beliefs to thermostats. A final word of common sense – while the differences between thermostats and human beings are, says Dennett, only a matter of degree, they are of such a degree that understanding the internal organisation of the one gives one very little basis for understanding that of the other.

### Why Does the Intentional Strategy Work?

- There are two very different answers to the ambiguous question of *why* the intentional strategy works as well as it does. The true but uninformative answer for simple intentional systems like thermostats is that they are designed to be systems that are easily comprehended and manipulated from the intentional stance. However, we really want to know what it is about the design that explains its performance – that is, how the machinery works.

- The same ambiguity arises if the system is a person. The first answer is that evolution has designed human beings to be rational, to believe and want what they ought to. Our long and demanding evolutionary ancestry makes the use of the intentional strategy a safe bet. While true and brief, this answer is also uninformative because what we want to know is how the machinery provided by Nature works. Unfortunately, we just don’t know the answer to the hard question, despite knowing the answer to the easy question and the fact that the strategy works.

- There’s no denying there are plenty of *doctrines* about. A Skinnerian behaviourist would say the strategy works because its imputations of beliefs and desires are shorthand for complex descriptions of prior histories of response and reinforcement. Saying someone desires ice-cream just *is* to say that previous
ingestion of ice-cream has been reinforced in him by the results, creating a propensity under further complex background conditions to engage in ice-cream-acquiring behaviour. Despite our lack of detailed knowledge of these historical facts we can still induct shrewd guesses, and these guesses are embodied in the claims of our intentional stance. Dennett thinks that, even were all this true, it would still tell us little about the way such propensities are regulated by the internal machinery.

- A more contemporary explanation is that the accounts of the workings of the strategy and mechanism will approximately coincide. For each predictively attributable belief there is a functionally relevant internal machine-state, decomposable into parts much as the sentence expressing the belief is into its words or terms. Inferences attributable to rational creatures are mirrored in physical, causal processes in the hardware, with logical form paralleled by the structural form of the corresponding states. This hypothesis is that there is a language of thought encoded in our brains, which will eventually be understood as symbol-manipulating machines analogous to computers. Dennett thinks this basic, bold claim of cognitive science will eventually prove correct.

- Dennett thinks that those who think it is obvious that such a theory will prove true are confusing two empirical claims. (1) The description provided by the intentional stance yields an objective, real pattern in the world, one missed by our Martians – this empirical claim is confirmed beyond sceptical doubt. (2) This real pattern is produced by another real and roughly isomorphic pattern within the brains of intelligent creatures – which can be doubted without doubting claim (1). Dennett thinks that, while there are reasons for believing (2), they are not overwhelming.

- Dennett accounts for the reasons as follows. As we progress from thermostat via robot to human being we encounter combinatorial explosion in our attempts to design systems. A 10% increase in inputs or more aspects of behaviour to control results in an increase in complexity of orders of magnitude. Things rapidly get out of hand and programs swamp the most powerful computers. Somehow the brain has solved the problem of combinatorial explosion. While a gigantic network of billions of cells, it remains finite, reliable, compact & swift – capable of learning new behaviours, vocabularies and theories almost without limit. Some generative, indefinitely extendable principles of representation must be involved, of which we have only one model – a human language. So, what else could we have but a language of thought? Our inability to think of alternatives warrants our pursuit of this strategy as far as we can, though we should remember that it is not guaranteed to be successful. One doesn’t well understand even a true empirical hypothesis if under the illusion that it is necessarily true.