2

True Believers: The Intentional Strategy and Why It Works

Death Speaks

There was a merchant in Baghdad who sent his servant to market to buy provisions and in a little while the servant came back, white and trembling, and said, Master, just now when I was in the market-place I was jostled by a woman in the crowd and when I turned I saw it was Death that jostled me. She looked at me and made a threatening gesture; now, lend me your horse, and I will ride away from this city and avoid my fate. I will go to Samarra and there Death will not find me. The merchant lent him his horse, and the servant mounted it, and he dug his spurs in its flanks and as fast as the horse could gallop he went. Then the merchant went down to the market-place and he saw me standing in the crowd, and he came to me and said, why did you make a threatening gesture to my servant when you saw him this morning? That was not a threatening gesture, I said, it was only a start of surprise. I was astonished to see him in Baghdad, for I had an appointment with him tonight in Samarra.

W. Somerset Maugham

In the social sciences, talk about *belief* is ubiquitous. Since social scientists are typically self-conscious about their methods, there is also a lot of talk about *talk about belief*. And since belief is a genuinely curious and perplexing phenomenon, showing many different faces to the world, there is abundant controversy. Sometimes belief attribution appears to be a dark, risky, and imponderable business—especially when exotic, and more particularly religious or superstitious, beliefs are in the limelight. These are not the only troublesome cases; we also

---

Originally presented as a Herbert Spencer Lecture at Oxford in November 1979, and reprinted, with permission, from A. F. Heath, ed., *Scientific Explanation* (Oxford: Oxford University Press, 1981).

court argument and skepticism when we attribute beliefs to nonhuman animals, or to infants, or to computers or robots. Or when the beliefs we feel constrained to attribute to an apparently healthy, adult member of our own society are contradictory, or even just wildly false. A biologist colleague of mine was once called on the telephone by a man in a bar who wanted him to settle a bet. The man asked: "Are rabbits birds?" "No" said the biologist. "Damn!" said the man as he hung up. Now could he *really* have believed that rabbits were birds? Could anyone really and truly be attributed that belief? Perhaps, but it would take a bit of a story to bring us to accept it.

In all of these cases belief attribution appears beset with subjectivity, infected with cultural relativism, prone to "indeterminacy of radical translation"—clearly an enterprise demanding special talents: the art of phenomenological analysis, hermeneutics, empathy, *Verstehen*, and all that. On other occasions, normal occasions, when familiar beliefs are the topic, belief attribution looks as easy as speaking prose and as objective and reliable as counting beans in a dish. Particularly when these straightforward cases are before us, it is quite plausible to suppose that in principle (if not yet in practice) it would be possible to confirm these simple, objective belief attributions by *finding something inside the believer's head*—by finding the beliefs themselves, in effect. "Look," someone might say, "You either believe there's milk in the fridge or you don't believe there's milk in the fridge" (you might have no opinion, in the latter case). But if you do believe this, that's a perfectly objective fact about you, and it must come down in the end to your brain's being in some particular physical state. If we knew more about physiological psychology, we could in principle determine the facts about your brain state and thereby determine whether or not you believe there is milk in the fridge, even if you were determined to be silent or disingenuous on the topic. In principle, on this view physiological psychology could trump the results—or nonresults—of any "black box" method in the social sciences that divines beliefs (and other mental features) by behavioral, cultural, social, historical, *external* criteria.

These differing reflections congeal into two opposing views on the nature of belief attribution, and hence on the nature of belief. The latter, a variety of *realism*, likens the question of whether a person has a particular belief to the question of whether a person is infected with a particular virus—a perfectly objective internal matter of fact about which an observer can often make educated guesses of great reliability. The former, which we could call *interpretationism* if we absolutely had to give it a name, likens the question of whether a person has a particular belief to the question of whether a person is immoral, or has style, or talent, or would make a good wife. Faced with such questions, we preface our answers with "well, it all depends on what you're interested in," or make some similar acknowledgment of the relativity of the issue. "It's a matter of interpretation," we say. These two opposing views, so baldly stated, do not fairly represent any serious theorists' positions, but they do express views that are typically seen as mutually exclusive and exhaustive; the theorist must be friendly with one and only one of these themes.

I think this is a mistake. My thesis will be that while belief is a perfectly objective phenomenon (that apparently makes me a realist), it can be discerned only from the point of view of one who adopts a certain *predictive strategy*, and its existence can be confirmed only by an assessment of the success of that strategy (that apparently makes me an interpretationist).

First I will describe the strategy, which I call the intentional strategy or adopting the intentional stance. To a first approximation, the intentional strategy consists of treating the object whose behavior you want to predict as a rational agent with beliefs and desires and other mental stages exhibiting what Brentano and others call *intentionality*. The strategy has often been described before, but I shall try to put this very familiar material in a new light by showing *how* it works and by showing *how well* it works.

Then I will argue that any object—or as I shall say, any *system*—whose behavior is well predicted by this strategy is in the fullest sense of the word a believer. *What it is* to be a true believer is to be an *intentional system*, a system whose behavior is reliably and voluminously predictable via the intentional strategy. I have argued for this position before (Dennett 1971, 1976b, 1978a), and my arguments have so far garnered few converts and many presumed counterexamples. I shall try again here, harder, and shall also deal with several compelling objections.

The Intentional Strategy and How It Works

There are many strategies, some good, some bad. Here is a strategy, for instance, for predicting the future behavior of a person: determine the date and hour of the person's birth and then feed this modest

datum into one or another astrological algorithm for generating predictions of the person's prospects. This strategy is deplorably popular. Its popularity is deplorable only because we have such good reasons for believing that it does not work (*pace* Feyerabend 1978). When astrological predictions come true this is sheer luck, or the result of such vagueness or ambiguity in the prophecy that almost any eventuality can be construed to confirm it. But suppose the astrological strategy did in fact work well on some people. We could call those people *astrological systems*—systems whose behavior was, as a matter of fact, predictable by the astrological strategy. If there were such people, such astrological systems, we would be more interested than most of us in fact are in *how the astrological strategy works*—that is, we would be interested in the rules, principles, or methods of astrology. We could find out how the strategy works by asking astrologers, reading their books, and observing them in action. But we would also be curious about *why* it worked. We might find that astrologers had no useful opinions about this latter question—they either had no theory of why it worked or their theories were pure hokum. Having a good strategy is one thing; knowing why it works is another.

So far as we know, however, the class of astrological systems is empty, so the astrological strategy is of interest only as a social curiosity. Other strategies have better credentials. Consider the physical strategy, or physical stance; if you want to predict the behavior of a system, determine its physical constitution (perhaps all the way down to the microphysical level) and the physical nature of the impingements upon it, and use your knowledge of the laws of physics to predict the outcome for any input. This is the grand and impractical strategy of Laplace for predicting the entire future of everything in the universe, but it has more modest, local, actually usable versions. The chemist or physicist in the laboratory can use this strategy to predict the behavior of exotic materials, but equally the cook in the kitchen can predict the effect of leaving the pot on the burner too long. The strategy is not always practically available, but that it will always work *in principle* is a dogma of the physical sciences (I ignore the minor complications raised by the subatomic indeterminacies of quantum physics.)

Sometimes, in any event, it is more effective to switch from the physical stance to what I call the design stance, where one ignores the actual (possibly messy) details of the physical constitution of an object, and, on the assumption that it has a certain design, predicts that

it will behave *as it is designed to behave* under various circumstances. For instance, most users of computers have not the foggiest idea what physical principles are responsible for the computer's highly reliable, and hence predictable, behavior. But if they have a good idea of what the computer is designed to do (a description of its operation at any one of the many possible levels of abstraction), they can predict its behavior with great accuracy and reliability, subject to disconfirmation only in cases of physical malfunction. Less dramatically, almost anyone can predict when an alarm clock will sound on the basis of the most casual inspection of its exterior. One does not know or care to know whether it is spring wound, battery driven, sunlight powered, made of brass wheels and jewel bearings or silicon chips—one just assumes that it is designed so that the alarm will sound when it is set to sound, and it is set to sound where it appears to be set to sound, and the clock will keep on running until that time and beyond, and is designed to run more or less accurately, and so forth. For more accurate and detailed design stance predictions of the alarm clock, one must descend to a less abstract level of description of its design; for instance, to the level at which gears are described, but their material is not specified.

Only the designed behavior of a system is predictable from the design stance, of course. If you want to predict the behavior of an alarm clock when it is pumped full of liquid helium, revert to the physical stance. Not just artifacts but also many biological objects (plants and animals, kidneys and hearts, stamens and pistils) behave in ways that can be predicted from the design stance. They are not just physical systems but designed systems.

Sometimes even the design stance is practically inaccessible, and then there is yet another stance or strategy one can adopt: the intentional stance. Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in many—but not all—instances yield a decision about what the agent ought to do; that is what you predict the agent *will* do.

The strategy becomes clearer with a little elaboration. Consider first how we go about populating each other's heads with beliefs. A few

truisms: sheltered people tend to be ignorant; if you expose someone to something he comes to know all about it. In general, it seems, we come to believe all the truths about the parts of the world around us we are put in a position to learn about. Exposure to x, that is, sensory confrontation with x over some suitable period of time, is the *normally sufficient* condition for knowing (or having true beliefs) about x. As we say, we come to *know all about* the things around us. Such exposure is only *normally* sufficient for knowledge, but this is not the large escape hatch it might appear; our threshold for accepting abnormal ignorance in the face of exposure is quite high. "I didn't know the gun was loaded," said by one who was observed to be present, sighted, and awake during the loading, meets with a variety of utter skepticism that only the most outlandish supporting tale could overwhelm.

Of course we do not come to learn or remember all the truths our sensory histories avail us. In spite of the phrase "know all about," what we come to know, normally, are only all the *relevant* truths our sensory histories avail us. I do not typically come to know the ratio of spectacle-wearing people to trousered people in a room I inhabit, though if this interested me, it would be readily learnable. It is not just that some facts about my environment are below my thresholds of discrimination or beyond the integration and holding power of my memory (such as the height in inches of all the people present), but that many perfectly detectable, graspable, memorable facts are of no interest to me and hence do not come to be believed by me. So one rule for attributing beliefs in the intentional strategy is this: attribute as beliefs all the truths relevant to the system's interests (or desires) that the system's experience to date has made available. This rule leads to attributing somewhat too much—since we all are somewhat forgetful, even of important things. It also fails to capture the false beliefs we are all known to have. But the attribution of false belief, *any* false belief, requires a special genealogy, which will be seen to consist in the main in true beliefs. Two paradigm cases: S believes (falsely) that p, because S believes (truly) that Jones told him that p, that Jones is pretty clever, that Jones did not intend to deceive him, . . . etc. Second case: S believes (falsely) that there is a snake on the barstool, because S believes (truly) that he seems to see a snake on the barstool, is himself sitting in a bar not a yard from the barstool he sees, and so forth. The falsehood has to start somewhere; the seed may be sown in hallucination, illusion, a normal variety of simple misperception, memory deterioration, or deliberate fraud, for instance, but the false beliefs that are reaped grow in a culture medium of true beliefs.

Then there are the arcane and sophisticated beliefs, true and false, that are so often at the focus of attention in discussions of belief attribution. They do not arise directly, goodness knows, from exposure to mundane things and events, but their attribution requires tracing out a lineage of mainly good argument or reasoning from the bulk of beliefs already attributed. An implication of the intentional strategy, then, is that true believers mainly believe truths. If anyone could devise an agreed-upon method of individuating and counting beliefs (which I doubt very much), we would see that all but the smallest portion (say, less than ten percent) of a person's beliefs were attributable under our first rule.[1]

Note that this rule is a derived rule, an elaboration and further

1. The idea that most of anyone's beliefs *must* be true seems obvious to some people. Support for the idea can be found in works by Quine, Putnam, Shoemaker, Davidson, and myself. Other people find the idea equally incredible—so probably each side is calling a different phenomenon belief. Once one makes the distinction between belief and opinion (in my technical sense—see "How to Change Your Mind" in *Brainstorms*, chapter 16), according to which opinions are linguistically infected, relatively sophisticated cognitive states—*roughly* states of betting on the truth of a particular, formulated sentence—one can see the near triviality of the claim that most beliefs are true. A few reflections on peripheral matters should bring it out. Consider Democritus, who had a systematic, all-embracing, but (let us say, for the sake of argument) entirely false physics. He had things *all wrong*, though his views held together and had a sort of systematic utility. But even if every *claim* that scholarship permits us to attribute to Democritus (either explicit or implicit in his writings) is false, these represent a vanishingly small fraction of his *beliefs*, which include both the vast numbers of humdrum standing beliefs he must have had (about which house he lived in, what to look for in a good pair of sandals, and so forth) and also those occasional beliefs that came and went by the millions as his perceptual experience changed.

But, it may be urged, this isolation of his humdrum beliefs from his science relies on an insupportable distinction between truths of observation and truths of theory; all Democritus's beliefs are theory-laden, and since his theory is false, they are false. The reply is as follows: Granted that all observation beliefs are theory laden, why should we choose Democritus's *explicit*, sophisticated theory (couched in his *opinions*) as the theory with which to burden his quotidian observations? Note that the least theoretical compatriot of Democritus also had myriads of theory-laden observation beliefs—and was, in one sense, none the wiser for it. Why should we not suppose Democritus's observations are laden with the same (presumably innocuous) theory? If Democritus forgot his theory, or changed his mind, his observational beliefs would be *largely* untouched. To the extent that his sophisticated theory played a discernible role in his routine behavior and expectations and so forth, it would be quite appropriate to couch his humdrum beliefs in terms of the sophisticated theory, but this will not yield a *mainly false* catalogue of beliefs, since so few of his beliefs will be affected. (The effect of theory on observation is nevertheless often underrated. See Churchland 1979 for dramatic and convincing examples of the tight relationship that can sometimes exist between theory and experience. [The discussion in this note was distilled from a useful conversation with Paul and Patricia Churchland and Michael Stack.])

specification of the fundamental rule: attribute those beliefs the system *ought to have*. Note also that the rule interacts with the attribution of desires. How do we attribute the desires (preferences, goals, interests) on whose basis we will shape the list of beliefs? We attribute the desires the system *ought to have*. That is the fundamental rule. It dictates, on a first pass, that we attribute the familiar list of highest, or most basic, desires to people: survival, absence of pain, food, comfort, procreation, entertainment. Citing any one of these desires typically terminates the "Why?" game of reason giving. One is not supposed to need an ulterior motive for desiring comfort or pleasure or the prolongation of one's existence. Derived rules of desire attribution interact with belief attributions. Trivially, we have the rule: attribute desires for those things a system believes to be good for it. Somewhat more informatively, attribute desires for those things a system believes to be best means to other ends it desires. The attribution of bizarre and detrimental desires thus requires, like the attribution of false beliefs, special stories.

The interaction between belief and desire becomes trickier when we consider what desires we attribute on the basis of verbal behavior. The capacity to *express* desires in language opens the floodgates of desire attribution. "I want a two-egg mushroom omelette, some French bread and butter, and a half bottle of lightly chilled white Burgundy." How could one begin to attribute a desire for anything so specific in the absence of such verbal declaration? How, indeed, could a creature come to *contract* such a specific desire without the aid of language? Language *enables* us to formulate highly specific desires, but it also *forces* us on occasion to commit ourselves to desires altogether more stringent in their conditions of satisfaction than anything we would otherwise have any reason to endeavor to satisfy. Since in order to get what you want you often have to say what you want, and since you often cannot say what you want without saying something more specific than you antecedently mean, you often end up giving others evidence—the very best of evidence, your unextorted word—that you desire things or states of affairs far more particular than would satisfy you—or better, than would have satisfied you, for once you have declared, being a man of your word, you acquire an interest in satisfying exactly the desire you declared and no other.

"I'd like some baked beans, please."

"Yes sir. How many?"

You might well object to having such a specification of desire demanded of you, but in fact we are all socialized to accede to similar requirements in daily life—to the point of not noticing it, and certainly not feeling oppressed by it. I dwell on this because it has a parallel in the realm of belief, where our linguistic environment is forever forcing us to give—or concede—precise verbal expression to convictions that lack the hard edges verbalization endows them with (see Dennett 1969, pp. 184–85, and *Brainstorms*, chapter 16). By concentrating on the *results* of this social force, while ignoring its distorting effect, one can easily be misled into thinking that it is *obvious* that beliefs and desires are rather like sentences stored in the head. Being language-using creatures, it is inevitable that we should often come to believe that some particular, actually formulated, spelled and punctuated sentence *is true*, and that on other occasions we should come to want such a sentence to *come true*, but these are special cases of belief and desire and as such may not be reliable models for the whole domain.

That is enough, on this occasion, about the principles of belief and desire attribution to be found in the intentional strategy. What about the rationality one attributes to an intentional system? One starts with the ideal of perfect rationality and revises downward as circumstances dictate. That is, one starts with the assumption that people believe all the implications of their beliefs and believe no contradictory pairs of beliefs. This does not create a practical problem of clutter (infinitely many implications, for instance), for one is interested only in ensuring that the system one is predicting is rational enough to get to the particular implications that are relevant to its behavioral predicament of the moment. Instances of irrationality, or of finitely powerful capacities of inferences, raise particularly knotty problems of interpretation, which I will set aside on this occasion (see chapter 4, "Making Sense of Ourselves," and Cherniak 1986).

For I want to turn from the description of the strategy to the question of its use. Do people actually use this strategy? Yes, all the time. There may someday be other strategies for attributing belief and desire and for predicting behavior, but this is the only one we all know now. And when does it work? It works with people almost all the time. Why would it *not* be a good idea to allow individual Oxford colleges to create and grant academic degrees whenever they saw fit? The answer is a long story, but very easy to generate. And there would be widespread agreement about the major points. We have no

difficulty thinking of the reasons people would then have for acting in such ways as to give others reasons for acting in such ways as to give others reasons for . . . creating a circumstance we would not want. Our use of the intentional strategy is so habitual and effortless that the role it plays in shaping our expectations about people is easily overlooked. The strategy also works on most other mammals most of the time. For instance, you can use it to design better traps to catch those mammals, by reasoning about what the creature knows or believes about various things, what it prefers, what it wants to avoid. The strategy works on birds, and on fish, and on reptiles, and on insects and spiders, and even on such lowly and unenterprising creatures as clams (once a clam believes there is danger about, it will not relax its grip on its closed shell until it is convinced that the danger has passed). It also works on some artifacts: the chess-playing computer will not take your knight because it knows that there is a line of ensuing play that would lead to losing its rook, and it does not want that to happen. More modestly, the thermostat will turn off the boiler as soon as it comes to believe the room has reached the desired temperature.

The strategy even works for plants. In a locale with late spring storms, you should plant apple varieties that are particularly *cautious* about *concluding* that it is spring—which is when they *want* to blossom, of course. It even works for such inanimate and apparently undesigned phenomena as lightning. An electrician once explained to me how he worked out how to protect my underground water pump from lightning damage: lightning, he said, always wants to find the best way to ground, but sometimes it gets tricked into taking second-best paths. You can protect the pump by making another, better path more *obvious* to the lightning.

True Believers as Intentional Systems

Now clearly this is a motley assortment of "serious" belief attributions, dubious belief attributions, pedagogically useful metaphors, *falçcons de parler*, and, perhaps worse, outright frauds. The next task would seem to be distinguishing those intentional systems that *really* have beliefs and desires from those we may find it handy to treat *as if* they had beliefs and desires. But that would be a Sisyphean labor, or else would be terminated by fiat. A better understanding of the phenomenon of belief begins with the observation that even in the worst

why Lectern is not an intentional system.

of these cases, even when we are surest that the strategy works *for the wrong reasons*, it is nevertheless true that it does work, at least a little bit. This is an interesting fact, which distinguishes this class of objects, the class of *intentional systems*, from the class of objects for which the strategy never works. But is this so? Does our definition of an intentional system exclude any objects at all? For instance, it seems the lectern in this lecture room can be construed as an intentional system, fully rational, believing that it is currently located at the center of the civilized world (as some of you may also think), and desiring above all else to remain at that center. What should such a rational agent so equipped with belief and desire do? Stay put, clearly, which is just what the lectern does. I predict the lectern's behavior, accurately, from the intentional stance, so is it an intentional system? If it is, anything at all is.

What should disqualify the lectern? For one thing, the strategy does not recommend itself in this case, for we get no predictive power from it that we did not antecedently have. We already knew what the lectern was going to do—namely nothing—and tailored the beliefs and desires to fit in a quite unprincipled way. In the case of people or animals or computers, however, the situation is different. In these cases often the only strategy that is at all practical is the intentional strategy; it gives us predictive power we can get by no other method. But, it will be urged, this is no difference in nature, but merely a difference that reflects upon our limited capacities as scientists. The Laplacean omniscient physicist could predict the behavior of a computer—or of a live human body, assuming it to be ultimately governed by the laws of physics—without any need for the risky, short-cut methods of either the design or intentional strategies. For people of limited mechanical aptitude, the intentional interpretation of a simple thermostat is a handy and largely innocuous crutch, but the engineers among us can quite fully grasp its internal operation without the aid of this anthropomorphizing. It may be true that the cleverest engineers find it practically impossible to maintain a clear conception of more complex systems, such as a time-sharing computer system or remote-controlled space probe, without lapsing into an intentional stance (and viewing these devices as asking and telling, trying and avoiding, wanting and believing), but this is just a more advanced case of human epistemic frailty. We would not want to classify these artifacts with the true believers—ourselves—on such variable and parochial grounds, would we? Would it not be intoler-

able to hold that some artifact or creature or person was a believer from the point of view of one observer, but not a believer at all from the point of view of another, cleverer observer? That would be a particularly radical version of interpretationism, and some have thought I espoused it in urging that belief be viewed in terms of the success of the intentional strategy. I must confess that my presentation of the view has sometimes invited that reading, but I now want to discourage it. The decision to adopt the intentional stance is free, but the facts about the success or failure of the stance, were one to adopt it, are perfectly objective.

Once the intentional strategy is in place, it is an extraordinarily powerful tool in prediction—a fact that is largely concealed by our typical concentration on the cases in which it yields dubious or unreliable results. Consider, for instance, predicting moves in a chess game. What makes chess an interesting game, one can see, is the *un*predictability of one's opponent's moves, except in those cases where moves are "forced"—where there is *clearly* one best move—typically the least of the available evils. But this unpredictability is put in context when one recognizes that in the typical chess situation there are very many perfectly legal and hence available moves, but only a few—perhaps half a dozen—with anything to be said for them, and hence only a few high-probability moves according to the intentional strategy. Even when the intentional strategy fails to distinguish a single move with a highest probability, it can dramatically reduce the number of live options.

The same feature is apparent when the intentional strategy is applied to "real world" cases. It is notoriously unable to predict the exact purchase and sell decisions of stock traders, for instance, or the exact sequence of words a politician will utter when making a scheduled speech, but one's confidence can be very high indeed about slightly less specific predictions: that the particular trader *will not buy utilities today*, or that the politician *will side with the unions against his party*, for example. This inability to predict fine-grained descriptions of actions, looked at another way, is a source of strength for the intentional strategy, for it is this neutrality with regard to details of implementation that permits one to exploit the intentional strategy in complex cases, for instance, in *chaining predictions* (see *Brainstorms*). Suppose the US Secretary of State were to announce he was a paid agent of the KGB. What an unparalleled event! How unpredictable its

consequences! Yet in fact we can predict dozens of not terribly interesting but perfectly salient consequences, and consequences of consequences. The President would confer with the rest of the Cabinet, which would support his decision to relieve the Secretary of State of his duties pending the results of various investigations, psychiatric and political, and all this would be reported at a news conference to people who would write stories that would be commented upon in editorials that would be read by people who would write letters to the editors, and so forth. None of that is daring prognostication, but note that it describes an arc of causation in space-time that could not be predicted under *any* description by any imaginable practical extension of physics or biology.

The power of the intentional strategy can be seen even more sharply with the aid of an objection first raised by Robert Nozick some years ago. Suppose, he suggested, some beings of vastly superior intelligence—from Mars, let us say—were to descend upon us, and suppose that we were to them as simple thermostats are to clever engineers. Suppose, that is, that they did not *need* the intentional stance—or even the design stance—to predict our behavior in all its detail. They can be supposed to be Laplacean super-physicists, capable of comprehending the activity on Wall Street, for instance, at the microphysical level. Where we see brokers and buildings and sell orders and bids, they see vast congeries of subatomic particles milling about—and they are such good physicists that they can predict days in advance what ink marks will appear each day on the paper tape labeled "Closing Dow Jones Industrial Average." They can predict the individual behaviors of all the various moving bodies they observe without ever treating any of them as intentional systems. Would we be right then to say that from *their* point of view we really were not believers at all (any more than a simple thermostat is)? If so, then our status as believers is nothing objective, but rather something in the eye of the beholder—provided the beholder shares our intellectual limitations.

Our imagined Martians might be able to predict the future of the human race by Laplacean methods, but if they did not also see us as intentional systems, they would be missing something perfectly objective: the *patterns* in human behavior that are describable from the intentional stance, and only from that stance, and that support generalizations and predictions. Take a particular instance in which

the Martians observe a stockbroker deciding to place an order for 500 shares of General Motors. They predict the exact motions of his fingers as he dials the phone and the exact vibrations of his vocal cords as he intones his order. But if the Martians do not see that indefinitely many *different* patterns of finger motions and vocal cord vibrations—even the motions of indefinitely many different individuals—could have been substituted for the actual particulars without perturbing the subsequent operation of the market, then they have failed to see a real pattern in the world they are observing. Just as there are indefinitely many ways of *being a spark plug*—and one has not understood what an internal combustion engine is unless one realizes that a variety of different devices can be screwed into these sockets without affecting the performance of the engine—so there are indefinitely many ways of *ordering 500 shares of General Motors*, and there are societal sockets in which one of these ways will produce just about the same effect as any other. There are also societal pivot points, as it were, where which way people go depends on whether they *believe that p*, or *desire A*, and does not depend on any of the other infinitely many ways they may be alike or different.

Suppose, pursuing our Martian fantasy a little further, that one of the Martians were to engage in a predicting contest with an Earthling. The Earthling and the Martian observe (and observe each other observing) a particular bit of local physical transaction. From the Earthling's point of view, this is what is observed. The telephone rings in Mrs. Gardner's kitchen. She answers, and this is what she says: "Oh, hello dear. You're coming home early? Within the hour? And bringing the boss to dinner? Pick up a bottle of wine on the way home, then, and drive carefully." On the basis of this observation, our Earthling predicts that a large metallic vehicle with rubber tires will come to a stop in the drive within one hour, disgorging two human beings, one of whom will be holding a paper bag containing a bottle containing an alcoholic fluid. The prediction is a bit risky, perhaps, but a good bet on all counts. The Martian makes the same prediction, but has to avail himself of much more information about an extraordinary number of interactions of which, so far as he can tell, the Earthling is entirely ignorant. For instance, the deceleration of the vehicle at intersection *A*, five miles from the house, without which there would have been a collision with another vehicle—whose collision course had been laboriously calculated over some hundreds of meters

by the Martian. The Earthling's performance would look like magic! How did the Earthling know that the human being who got out of the car and got the bottle in the shop would get back in? The coming true of the Earthling's prediction, after all the vagaries, intersections, and branches in the paths charted by the Martian, would seem to anyone bereft of the intentional strategy as marvelous and inexplicable as the fatalistic inevitability of the appointment in Samarra. Fatalists—for instance, astrologers—believe that there is a pattern in human affairs that is inexorable, that will impose itself *come what may*, that is, no matter how the victims scheme and second-guess, no matter how they twist and turn in their chains. These fatalists are wrong, but they are *almost* right. There *are* patterns in human affairs that impose themselves, not quite inexorably but with great vigor, absorbing physical perturbations and variations that might as well be considered random; these are the patterns that we characterize in terms of the beliefs, desires, and intentions of rational agents.

No doubt you will have noticed, and been distracted by, a serious flaw in our thought experiment: the Martian is presumed to treat his Earthling opponent as an intelligent being like himself, with whom communication is possible, a being with whom one can make a wager, against whom one can compete. In short, a being with beliefs (such as the belief he expressed in his prediction) and desires (such as the desire to win the prediction contest). So if the Martian sees the pattern in one Earthling, how can he fail to see it in the others? As a bit of narrative, our example could be strengthened by supposing that our Earthling cleverly learned Martian (which is transmitted by X-ray modulation) and disguised himself as a Martian, counting on the species-chauvinism of these otherwise brilliant aliens to permit him to pass as an intentional system while not giving away the secret of his fellow human beings. This addition might get us over a bad twist in the tale, but might obscure the moral to be drawn: namely, *the unavoidability of the intentional stance with regard to oneself and one's fellow intelligent beings*. This unavoidability is itself interest relative; it is perfectly possible to adopt a physical stance, for instance, with regard to an intelligent being, oneself included, but not to the exclusion of maintaining at the same time an intentional stance with regard to oneself at a minimum, and one's fellows *if* one intends, for instance, to learn what they know (a point that has been powerfully made by Stuart Hampshire in a number of writings). We can perhaps suppose

our super-intelligent Martians fail to recognize *us* as intentional systems, but we cannot suppose them to lack the requisite concepts.[2] If they observe, theorize, predict, communicate, they view *themselves* as intentional systems.[3] Where there are intelligent beings, the patterns must be there to be described, whether or not we care to see them.

It is important to recognize the objective reality of the intentional patterns discernible in the activities of intelligent creatures, but also important to recognize the incompleteness and imperfections in the patterns. The objective fact is that the intentional strategy *works as well as it does*, which is not perfectly. No one is perfectly rational, perfectly unforgetful, all-observant, or invulnerable to fatigue, malfunction, or design imperfection. This leads inevitably to circumstances beyond the power of the intentional strategy to describe, in much the same way that physical damage to an artifact, such as a telephone or an automobile, may render it indescribable by the normal design terminology for that artifact. How do you draw the schematic wiring diagram of an audio amplifier that has been partially melted, or how do you characterize the program state of a malfunctioning computer? In cases of even the mildest and most familiar cognitive pathology—where people seem to hold contradictory beliefs or to be deceiving themselves, for instance—the canons of interpretation of the intentional strategy fail to yield clear, stable verdicts about which beliefs and desires to attribute to a person.

Now a *strong* realist position on beliefs and desires would claim that in these cases the person in question really does have some particular beliefs and desires which the intentional strategy, as I have described it, is simply unable to divine. On the milder sort of realism I am advocating, there is no fact of the matter of exactly which beliefs and desires a person has in these degenerate cases, but this is not a surrender to relativism or subjectivism, for *when* and *why* there is no fact of the matter is itself a matter of objective fact. On this view one can even acknowledge the *interest relativity* of belief attributions and grant that given the different interests of different cultures, for instance, the beliefs and desires one culture would attribute to a member might be quite different from the beliefs and desires another culture would attribute to that very same person. But supposing that were so in a particular case, there would be the further facts about *how well* each of the rival intentional strategies worked for predicting the behavior of that person. We can be sure in advance that no intentional interpretation of an individual will work to perfection, and it may be that two rival schemes are about equally good, and better than any others we can devise. That this is the case is itself something about which there can be a fact of the matter. The objective presence of one pattern (with whatever imperfections) does not rule out the objective presence of another pattern (with whatever imperfections).

2. A member of the audience in Oxford pointed out that if the Martian included the Earthling in his physical stance purview (a possibility I had not explicitly excluded), he would not be surprised by the Earthling's prediction. He would indeed have predicted exactly the pattern of X-ray modulations produced by the Earthling speaking Martian. True, but as the Martian wrote down the results of his calculations, his prediction of the Earthling's prediction would appear, word by Martian word, as on a Ouija board, and what would be baffling to the Martian was how this chunk of mechanism, the Earthling predictor dressed up like a Martian, was able to yield this *true* sentence of Martian when it was so informationally isolated from the events the Martian needed to know of in order to make his own prediction about the arriving automobile.

3. Might there not be intelligent beings who had no use for communicating, predicting, observing . . . ? There might be marvelous, nifty, invulnerable entities lacking these modes of action, but I cannot see what would lead us to call them *intelligent*.

The bogey of radically different interpretations with equal warrant from the intentional strategy is theoretically important—one might better say metaphysically important—but practically negligible once one restricts one's attention to the largest and most complex intentional systems we know: human beings.[4]

Until now I have been stressing our kinship to clams and thermostats, in order to emphasize a view of the logical status of belief attribution, but the time has come to acknowledge the obvious differences and say what can be made of them. The perverse claim remains: *all there is* to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy, and hence *all there is* to really and truly believing that p (for any proposition p) is being an intentional system for which p occurs as a belief in the best (most predictive) interpretation. But once we turn our attention to the truly interesting and versatile intentional systems, we see that this apparently shallow and instrumentalistic criterion of belief puts a severe constraint on the internal constitution of a genuine believer, and thus yields a robust version of belief after all.

Consider the lowly thermostat, as degenerate a case of an inten-

4. John McCarthy's analogy to cryptography nicely makes this point. The larger the corpus of cipher text, the less chance there is of dual, systematically unrelated decipherings. For a very useful discussion of the principles and presuppositions of the intentional stance applied to machines—explicitly including thermostats—see McCarthy 1979.

tional system as could conceivably hold our attention for more than a moment. Going along with the gag, we might agree to grant it the capacity for about half a dozen different beliefs and fewer desires—it can believe the room is too cold or too hot, that the boiler is on or off, and that if it wants the room warmer it should turn on the boiler, and so forth. But surely this is imputing too much to the thermostat; it has no concept of heat or of a boiler, for instance. So suppose we *de-interpret* its beliefs and desires: it can believe the *A* is too *F* or *G*, and if it wants the *A* to be more *F* it should do *K*, and so forth. After all, by attaching the thermostatic control mechanism to different input and output devices, it could be made to regulate the amount of water in a tank, or the speed of a train, for instance. Its attachment to a heat-sensitive transducer and a boiler is too impoverished a link to the world to grant any rich semantics to its belief-like states.

But suppose we then enrich these modes of attachment. Suppose we give it more than one way of learning about the temperature, for instance. We give it an eye of sorts that can distinguish huddled, shivering occupants of the room and an ear so that it can be told how cold it is. We give it some facts about geography so that it can conclude that it is probably in a cold place if it learns that its spatio-temporal location is Winnipeg in December. Of course giving it a visual system that is multipurpose and general—not a mere shivering-object detector—will require vast complications of its inner structure. Suppose we also give our system more behavioral versatility: it chooses the boiler fuel, purchases it from the cheapest and most reliable dealer, checks the weather stripping, and so forth. This adds another dimension of internal complexity; it gives individual belief-like states *more to do*, in effect, by providing more and different occasions for their derivation or deduction from other states, and by providing more and different occasions for them to serve as premises for further reasoning. The cumulative effect of enriching these connections between the device and the world in which it resides is to enrich the semantics of its dummy predicates, *F* and *G* and the rest. The more of this we add, the less amenable our device becomes to serving as the control structure of anything other than a room-temperature maintenance system. A more formal way of saying this is that the class of indistinguishably satisfactory models of the formal system embodied in its internal states gets smaller and smaller as we add such complexities; the more we add, the richer or more demanding or specific the semantics of the system, until eventually we reach systems for which a unique semantic interpretation is practically (but never in principle) dictated (cf. Hayes 1979). At that point we say this device (or animal or person) has beliefs *about heat* and *about this very room*, and so forth, not only because of the system's actual location in, and operations on, the world, but because we cannot imagine another niche in which it could be placed *where it would work* (see also chapters 5 and 8).

Our original simple thermostat had a state we called a belief about a particular boiler, to the effect that it was on or off. Why about *that* boiler? Well, what other boiler would you want to say it was about? The belief is about the boiler because it is *fastened* to the boiler.[5] Given the actual, if mimimal, causal link to the world that happened to be in effect, we could endow a state of the device with *meaning* (of a sort) and *truth conditions*, but it was altogether too easy to substitute a different minimal link and completely change the meaning (in this impoverished sense) of that internal state. But as systems become perceptually richer and behaviorally more versatile, it becomes harder and harder to make substitutions in the actual links of the system to the world without changing the organization of the system itself. If you change its environment, it will *notice*, in effect, and make a change in its internal state in response. There comes to be a two-way constraint of growing specificity between the device and the environment. Fix the device in any one state and it demands a very specific environment in which to operate properly (you can no longer switch it easily from regulating temperature to regulating speed or anything else); but at the same time, if you do not *fix* the state it is in, but just plonk it down in a changed environment, its sensory attachments will be sensitive and discriminative enough to respond appropriately to the change, driving the system into a new state, in which it will operate effectively in the new environment. There is a familiar way of alluding to this tight relationship that can exist between the organization of a system and its environment: you say that the organism continuously *mirrors* the environment, or that there is a *representation* of the environment in—or implicit in—the organization of the system.

5. This idea is the ancestor in effect of the species of different ideas lumped together under the rubric of *de re* belief. If one builds from this idea toward its scions, one can see better the difficulties with them, and how to repair them. (For more on this topic, see chapter 5, "Beyond Belief.")

It is not that we attribute (or should attribute) beliefs and desires only to things in which we find internal representations, but rather that when we discover some object for which the intentional strategy works, we endeavor to interpret some of its internal states or processes as internal representations. What makes some internal feature of a thing a representation could only be its role in regulating the behavior of an intentional system.

Now the reason for stressing our kinship with the thermostat should be clear. There is no magic moment in the transition from a simple thermostat to a system that *really* has an internal representation of the world around it. The thermostat has a minimally demanding representation of the world, fancier thermostats have more demanding representations of the world, fancier robots for helping around the house would have still more demanding representations of the world. Finally you reach us. We are so multifariously and intricately connected to the world that almost no substitution is possible—though it is clearly imaginable in a thought experiment. Hilary Putnam imagines the planet Twin Earth, which is just like Earth right down to the scuff marks on the shoes of the Twin Earth replica of your neighbor, but which differs from Earth in some property that is entirely beneath the thresholds of your capacities to discriminate. (What they call water on Twin Earth has a different chemical analysis.) Were *you* to be whisked instantaneously to Twin Earth and exchanged for your Twin Earth replica, you would never be the wiser—just like the simple control system that cannot tell whether it is regulating temperature, speed, or volume of water in a tank. It is easy to devise radically different Twin Earths for something as simple and sensorily deprived as a thermostat, but your internal organization puts a much more stringent demand on substitution. Your Twin Earth and Earth must be virtual replicas or you will change state dramatically on arrival.

So which boiler are *your* beliefs about when you believe the boiler is on? Why, the boiler in your cellar (rather than its twin on Twin Earth, for instance). What other boiler would your beliefs be about? The completion of the semantic interpretation of your beliefs, fixing the referents of your beliefs, requires, as in the case of the thermostat, facts about your actual embedding in the world. The principles, and problems, of interpretation that we discover when we attribute beliefs to people are the *same* principles and problems we discover when we look at the ludicrous, but blessedly simple, problem of attributing beliefs to a thermostat. The differences are of degree, but nevertheless of such great degree that understanding the internal organization of a simple intentional system gives one very little basis for understanding the internal organization of a complex intentional system, such as a human being.

Why Does the Intentional Strategy Work?

When we turn to the question of *why* the intentional strategy works as well as it does, we find that the question is ambiguous, admitting of two very different sorts of answers. If the intentional system is a simple thermostat, one answer is simply this: the intentional strategy works because the thermostat is well designed; it was designed to be a system that could be easily and reliably comprehended and manipulated from this stance. That is true, but not very informative, if what we are after are the actual features of its design that explain its performance. Fortunately, however, in the case of a simple thermostat those features are easily discovered and understood, so the other answer to our *why* question, which is really an answer about *how the machinery works*, is readily available.

If the intentional system in question is a person, there is also an ambiguity in our question. The first answer to the question of why the intentional strategy works is that evolution has designed human beings to be rational, to believe what they ought to believe and want what they ought to want. The fact that we are products of a long and demanding evolutionary process guarantees that using the intentional strategy on us is a safe bet. This answer has the virtues of truth and brevity, and on this occasion the additional virtue of being an answer Herbert Spencer would applaud, but it is also strikingly uninformative. The more difficult version of the question asks, in effect, how the machinery which Nature has provided us works. And we cannot yet give a good answer to that question. We just do not know. We do know how the *strategy* works, and we know the easy answer to the question of why it works, but knowing these does not help us much with the hard answer.

It is not that there is any dearth of doctrine, however. A Skinnerian behaviorist, for instance, would say that the strategy works because its imputations of beliefs and desires are shorthand, in effect, for as yet unimaginably complex descriptions of the effects of prior histories of response and reinforcement. To say that someone wants some ice cream is to say that in the past the ingestion of ice cream has been

reinforced in him by the results, creating a propensity under certain background conditions (also too complex to describe) to engage in ice-cream-acquiring behavior. In the absence of detailed knowledge of those historical facts we can nevertheless make shrewd guesses on inductive grounds; these guesses are embodied in our intentional stance claims. Even if all this were true, it would tell us very little about the way such propensities were regulated by the internal machinery.

A currently more popular explanation is that the account of how the strategy works and the account of how the mechanism works will (roughly) *coincide:* for each predictively attributable belief, there will be a functionally salient internal state of the machinery, decomposable into functional parts in just about the same way the sentence expressing the belief is decomposable into parts—that is, words or terms. The inferences we attribute to rational creatures will be mirrored by physical, causal processes in the hardware; the *logical* form of the propositions believed will be copied in the *structural* form of the states in correspondence with them. This is the hypothesis that there is a *language of thought* coded in our brains, and our brains will eventually be understood as symbol manipulating systems in at least rough analogy with computers. Many different versions of this view are currently being explored, in the new research program called cognitive science, and provided one allows great latitude for attenuation of the basic, bold claim, I think some version of it will prove correct.

But I do not believe that this is *obvious*. Those who think that it is obvious, or inevitable, that such a theory will prove true (and there are many who do), are confusing two different empirical claims. The first is that intentional stance description yields an objective, real pattern in the world—the pattern our imaginary Martians missed. This is an empirical claim, but one that is confirmed beyond skepticism. The second is that this real pattern is *produced by* another real pattern roughly isomorphic to it within the brains of intelligent creatures. Doubting the existence of the second real pattern is not doubting the existence of the first. There *are* reasons for believing in the second pattern, but they are not overwhelming. The best simple account I can give of the reasons is as follows.

As we ascend the scale of complexity from simple thermostat, through sophisticated robot, to human being, we discover that our efforts to design systems with the requisite behavior increasingly run

foul of the problem of *combinatorial explosion*. Increasing some parameter by, say, ten percent—ten percent more inputs or more degrees of freedom in the behavior to be controlled or more words to be recognized or whatever—tends to increase the internal complexity of the system being designed by orders of magnitude. Things get out of hand very fast and, for instance, can lead to computer programs that will swamp the largest, fastest machines. Now somehow the brain has solved the problem of combinatorial explosion. It is a gigantic network of billions of cells, but still finite, compact, reliable, and swift, and capable of learning new behaviors, vocabularies, theories, almost without limit. Some elegant, *generative*, indefinitely extendable principles of representation must be responsible. We have only one model of such a representation system: a human language. So the argument for a language of thought comes down to this: what else could it be? We have so far been unable to imagine any plausible alternative in any detail. That is a good enough reason, I think, for recommending as a matter of scientific tactics that we pursue the hypothesis in its various forms as far as we can.[6] But we will engage in that exploration more circumspectly, and fruitfully, if we bear in mind that its inevitable rightness is far from assured. One does not well understand even a true empirical hypothesis so long as one is under the misapprehension that it is necessarily true.

6. The fact that all *language of thought* models of mental representation so far proposed fall victim to combinatorial explosion in one way or another should temper one's enthusiasm for engaging in what Fodor aptly calls "the only game in town."