5

DENNETT ON INTENTIONAL SYSTEMS

DURING THE LAST dozen years, Daniel Dennett has been elaborating an interconnected—and increasingly influential—set of views in the philosophy of mind, the philosophy of psychology, and those parts of moral philosophy that deal with the notions of freedom, responsibility and personhood. The central unifying theme running through Dennett's writings on each of these topics is his concept of an *intentional system*. He invokes the concept to "legitimize" mentalistic predicates (*Brainstorms*, p. xvii),[1] to explain the theoretical strategy of cognitive psychology and artificial intelligence, and, ultimately, to attempt a reconciliation between "our vision of ourselves as responsible, free, rational agents, and our vision of ourselves as complex parts of the physical world of science" (*BS*, p. x). My goal in this paper is to raise some doubts about the "intentional coin" (*BS*, p. xviii) with which Dennett proposes to purchase his moral and "mental treasures." Since I aim to offer a critique of Dennett's views, it is inevitable that much of what I say will be negative in tone. But this tone should not be misconstrued. It is my view that Dennett's

1. References to Dennett's writings will be identified in parentheses in the text. I will use the following abbreviations:

BS = Daniel Dennett, *Brainstorms* (Montgomery, VT: Bradford Books, 1978).
TK = Daniel Dennett, "Three kinds of intentional psychology," in *Reduction, Time, and Reality* (ed. R. A. Healey), Cambridge University Press, 1981.
R = Daniel Dennett, "Reply to Professor Stich," *Philosophical Books*, 21, 2 (April, 1980).
TB = Daniel Dennett, "True believers: The intentional strategy and why it works," reprinted in this volume above.

theories are of great importance and will shape discussion in the philosophy of mind for decades to come. Moreover, I think that much of what Dennett says is close to being true. If we reconstruct his notion of an intentional system to eliminate its instrumentalism and its unfortunate infatuation with idealized rationality, we can use the result to give a better account of commonsense mentalistic notions, and also to give a clearer and more tenable account of the strategy of cognitive science. Toward the end of this paper I will sketch the outlines of such a "derationalized" cousin to Dennett's idea of an intentional system.

I

In explaining the idea of an intentional system, Dennett's recurrent illustration is the chess-playing computer. There are, he urges, three quite different stances we might "adopt in trying to predict and explain its behavior" (*BS*, p. 237).

> First there is the *design stance*. If one knows exactly how the computer's program has been designed . . . one can predict the computer's designed response to any move one makes. One's prediction will come true provided only that the computer performs as designed, that is, without breakdown. . . . The essential feature of the design stance is that we make predictions solely from knowledge or assumptions about the system's design, often without making any examination of the innards of the particular object.
>
> Second, there is what we may call the *physical stance*. From this stance our predictions are based on the actual state of the particular system, and are worked out by applying whatever knowledge we have of the laws of nature. . . . One seldom adopts the physical stance in dealing with a computer just because the number of critical variables in the physical constitution of a computer would overwhelm the most prodigious human calculator. . . . Attempting to give a physical account or prediction of the chess playing computer would be a pointless and herculean labor, but it would work in principle. One could predict the response it would make in a chess game by tracing out the effects of the input energies all the way through the computer until once more type was pressed against paper and a response was printed.
>
> There is a third stance one can adopt toward a system, and that is the *intentional stance*. This tends to be the most appropriate when the system one is dealing with is too complex to be dealt with effectively from the other stances. In the case of a chess playing computer one adopts this stance when one tries to predict its response to one's move by figuring out what a good or reasonable response would be, given the information the computer has about the situation. Here one assumes not just the absence of malfunction but the rationality of the design or programming as well.
>
> Whenever one can successfully adopt the intentional stance toward an object, I call that object an *intentional* system. The success of the stance is of course a matter

> settled pragmatically, without reference to whether the object *really* has beliefs, intentions, and so forth; so whether or not any computer can be conscious, or have thoughts or desires, some computers undeniably *are* intentional systems, for they are systems whose behavior can be predicted, and most efficiently predicted, by adopting the intentional stance towards them. (*BS*, pp. 237–8; for a largely identical passage, cf. *BS*, pp, 4–7.)

So *any* object will count as an intentional system if we can usefully predict its behavior by assuming that it will behave *rationally*. And what is it to behave rationally? Here, Dennett suggests, the full answer must ultimately be provided by a new sort of theory, *intentional-system theory*, which will provide us with a *normative* account of rationality. This new theory "is envisaged as a close kin of—and overlapping with—such already existing disciplines as epistemic logic, decision theory and game theory, which are all similarly abstract, normative and couched in intentional language" (*TK*, p. 19). Of course, we already have some "rough and ready principles" of rationality which we can and do press into service pending a more detailed normative theory:

1. A system's beliefs are those it *ought to have*, given its perceptual capacities, its epistemic needs, and its biography. Thus in general, its beliefs are both true and relevant to its life. . . .
2. A system's desires are those it ought to have, given its biological needs and the most practicable means of satisfying them. Thus [naturally evolved] intentional systems desire survival and procreation, and hence desire food, security, health, sex, wealth, power, influence, and so forth, and also whatever local arrangements tend (in their eyes—given their beliefs) to further these ends in appropriate measure. . . .
3. A system's behavior will consist of those acts that *it would be rational* for an agent with those beliefs and desires to perform. (*TK*, pp. 8–9)

Obviously these three principles are very rough and ready indeed. However, we also have a wealth of more detailed common-sense principles that anchor our intuitive notion of rationality. Some of these, in turn, are systematized and improved upon by existing theories in logic, evolutionary biology and decision theory. But though the intentional-system theorist can count on some help from these more developed disciplines, he still has a great deal of work to do. Neither singly nor severally do these disciplines tell us what beliefs a given organism or system ought to have, what desires it ought to have, or how it should act, given the beliefs and desires it has. Dennett has no illusions on the point. He portrays intentional-system theory—the general normative theory of rationality—as a discipline in its infancy. When the course of our argument requires some substantive premises about what it would be rational for a system to believe or do, we can follow Dennett's lead and let our common-sense intuitions be our guide.

I have been stressing the role of a normative theory of rationality in Dennett's account of the intentional stance. But there is a second, equally important, component in his view. According to Dennett, when we describe an organism or an artifact as an intentional system, we are making no commitments about the internal physical workings of the system. *Nor are we saying anything about the design or program of the system*. Just as a single program or design description is compatible with indefinitely many physical realizations, so too a single intentional description is compatible with indefinitely many different programs or design descriptions. To view an object as an intentional system we must attribute to it a substantial range of beliefs and desires—the beliefs and desires it would be rational for such an object to have, given its nature and history. However, we need not assume that the beliefs and desires attributed correspond in any systematic way to internal states characterized either physically or functionally. Dennett makes the point vividly with the example of two robots each designed to be identical to a given person, Mary, when viewed from the intentional stance. The first robot, Ruth, "has internal processes which 'model' Mary's as closely as you like" (*BS*, p. 105). It is functionally identical to Mary, though the two may be quite different physically. Since Mary and Ruth share a common design or program, they will behave identically. Thus any beliefs and desires we attribute to Mary we may attribute also to Ruth, and the attributions will be equally useful in predicting their behavior. The second robot, Sally, has a program which is input–output equivalent to Ruth's, though it uses a quite different computational strategy. "Sally may not be a very good psychological model of Mary," since "Sally's response delays, errors and the like may not match Mary's." But at the level of common-sense descriptions of actions, all three will behave alike. ". . . the ascription of all Mary's beliefs and desires (etc.) to Sally will be just as predictive as their ascription to Ruth so far as prediction of action goes" (*BS*, p. 105). So when we adopt the intentional stance, Mary, Ruth and Sally are indistinguishable.

Dennett, then, is a self-professed instrumentalist about the beliefs and desires we ascribe to an object when we adopt the intentional stance toward it. ". . . the beliefs and other intentions of an intentional systems need [not] be *represented* 'within' the system in any way for us to get a purchase on predicting its behavior by ascribing such intentions to it" (*BS*, p. 277). Rather, these "putative . . . states" can be relegated "to the role of idealized fictions in an action-predicting, action-explaining calculus" (*BS*, p. 30). For Dennett, the belief and desire states of an intentional system are not what Reichenbach calls "illata—posited theoretical entities." Rather they are "abstracta—calculation bound entities or logical constructs" (*TK*, p. 20). Of course, it is conceivable that some objects which are usefully treated as intentional systems really do have internal states that correspond to the beliefs and desires ascribed to them in an intentional characterization. As some writers have suggested, there might be functionally distinct neural belief and desire stores where each belief and desire is inscribed in an appropriate neural code. Dennett, however, thinks this is not likely to be true for people, animals and other familiar intentional

systems.[2] Be this as it may, the important point in the present context is that when we describe an object in intentional-system terms, we are quite explicitly *not* making any commitment about its workings, beyond the minimal claim that whatever the mechanism causally responsible for the behavior may be, it must be the sort of mechanism which will produce behavior generally predictable by assuming the intentional stance.

This completes my sketch of Dennett's notion of intentional systems. Let us now consider what Dennett wants to do with the notion. The principal project Dennett has in mind for intentional systems is "legitimizing" (*BS*, p. xvii), or providing a sort of "conceptual reduction" (*TK*, p. 30) of various notions in common-sense or folk psycology. The sort of legitimizing Dennett has in mind is explained by analogy with Church's Thesis. Church proposed that the informal, intuitive mathematical concept of an "effective" procedure be identified with the formal notion of a recursive (or Turing-machine computable) function. The proposal "is not provable, since it hinges on the intuitive and unformalizable notion of an effective procedure, but it is generally accepted, and it provides a very useful reduction of a fuzzy-but-useful mathematical notion to a crisply defined notion of apparently equal scope and greater power" (*BS*, p. xviii; cf. also *TK*, p. 30). It is Dennett's hope to provide the same sort of legitimization of the notions of folk psychology by showing how these notions can be characterized in terms of the notions of intentional-system theory. ". . . the claim that every mental phenomenon alluded to in folk psychology is *intentional-system-characterizable* would, if true, provide a reduction of the mental as ordinarily understood—a domain whose boundaries are at best fixed by mutual acknowledgement and shared intuition—to a clearly defined domain of entities, whose principles of organization are familiar, relatively formal and systematic, and entirely general" (*TK*, pp. 30–1).

All this sounds reasonable enough—an exicting project, if Dennett can pull it off. The effort looks even more intriguing when we note how broadly Dennett intends to cast his net. It is his aim to show not only that such "program receptive" (*BS*, p. 29) features of mentality as belief and desire are intentional-system-characterizable, but also that "program resistant features of mentality" like pain, dreams, mental images, and even free will are "captured in the net of intentional systems" (*BS*, p. xviii). But a dark cloud looms on the horizon, one that will continue to plague us. In much of his work Dennett exhibits an exasperating tendency to make bold, flamboyant, fascinating claims in one breath, only to take them back, or seem to, in the next. Thus, scarcely a page after proclaiming his intention to show that a broad range of common-sense mental phenomena are intentional-system-characterizable and thus legitimized, Dennett proclaims himself to be an eliminative materialist concerning these very same phenomena. Beliefs, desires, pains, mental images, experiences—as these are ordinarily understood—"are not good

2. For his arguments on this point, cf. "Brain writing and mind reading," (*BS*, pp. 39–50) and "A cure for the common code," (*BS*, pp. 90–108).

theoretical entities, however well entrenched" (*BS*, p. xx) the terms 'belief,' 'pain,' etc. may be in the habits of thought of our society. So "we legislate the putative items right out of existence" (*BS*, p. xx). How are we to make sense of this apparent contradiction?

There is, I think, a plausible—and uncontradictory—interpretation of what Dennett is up to. The problem he is grappling with is that the fit between our intuitive folk-psychological notions and the intentional-system characterizations he provides for them is just not as comfortable as the fit between the intuitive notion of effective mathematical procedure and the formal notion of Turing computability. Our folk-psychological concepts, "like folk productions generally," are complex, messy, variegated and in danger of incoherence (*TK*, p. 16). By contrast, notions characterized in terms of intentional-system theory are—it is to be hoped—coherent, sharply drawn and constructed with a self-conscious eye for their subsequent incorporation into science (*TK*, p. 6). The intentional-system analysans are intended to be improvements on their analysanda. What they give us is not an "anthropological" (*TK*, p. 6) portrait of our folk notions (warts and all), but rather an improved version of "the parts of folk psychology worth caring about" (*TK*, p. 30). So Dennett is an eliminative materialist about mental phenomena alluded to in warts-and-all folk psychology; what are intentional-system-characterizable are not the notions of folk psychology, but rather related successor concepts which capture all that's worth caring about.

But now what are we to make of the claim that the intentional system *Ersätze* capture all that's worth caring about in folk psychology: What *is* worth caring about? Dennett concedes that an "anthropological" study of unreconstructed folk notions which includes "whatever folk actually include in their theory, however misguided, incoherent, gratuitous some of it may be," (*TK*, p. 6) would be a perfectly legitimate endeavor. Folk theory may be myth, "but it is a myth we live in, so it is an 'important' phenomenon in nature" (*TK*, p. 6).[3] However, Dennett does not share the anthropologist's (or the cognitive simulator's) interest in the idiosyncrasies and contradictions embedded in our folk notions. What is of interest to him, he strongly suggests, is "the proto-scientific quest": "an attempt to prepare folk theory for subsequent incorporation into or reduction to the rest of science," eliminating "all that is false or ill-founded" (*TK*, p. 6). If matters stopped there, we could parse Dennett's "all that's worth caring about" as "all that's worth caring about for the purposes of science." But matters do not stop there. To see why, we will have to take a detour to survey another central theme in Dennett's thinking.

As we have noted, a basic goal of Dennett's theory is to reconcile "our vision of ourselves as responsible, free, rational agents, and our vision of ourselves as complex parts

3. This "anthropological quest," when pursued systematically is the business of the cognitive simulator. Cf., for example, Roger Shank and Robert Abelson, *Scripts, Plans, Goals and Understanding* (Hillsdale, NJ: Lawrence Erlbaum Associates, 1977); also Aaron Slomon, *The Computer Revolution in Philosophy* (Atlantic Highlands, NJ: Humanities Press, 1978), ch. 4.

of the physical world of science" (*BS*, p. x). The conflict that threatens between these two visions is a perennial philosophical preoccupation:

> the validity of our conceptual scheme of moral agents having dignity, freedom and responsibility stands or falls on the question: can men ever be truly said to have beliefs, desires, intentions? If they can, there is at least some hope of retaining a notion of the dignity of man; if they cannot, if men never can be said truly to want or believe, then surely they never can be said truly to act responsibly, or to have a conception of justice, or to know the difference between right and wrong. (*BS*, pp. 63–4)

Yet many psychologists, most notoriously Skinner, have denied that people have beliefs, desires and other mental states.[4] This threat to our view of ourselves as moral agents does not arise only from rabid behaviorism. Dennett sees it lurking also in certain recently fashionable philosophical theories about the nature of mental states. Consider, for example, the type-type identity theory which holds that every mental-state type is to be identified with a physical-state type—a brain state characterized in physico-chemical terms. What if it should turn out that there simply is *no* physical-state type that is shared by all beings to whom we commonly attribute the belief that snow is white? If we hang on to the type-type identity theory, then this very plausible empirical finding would seem to entail that there is no such mental state as believing that snow is white. Much the same result threatens from those versions of functionalism which hold that "each mental type is identifiable as a functional type in the language of Turing machine description" (*BS*, p. xvi). For "there is really no more reason to believe you and I 'have the same program' in *any* relaxed and abstract sense, considering the differences in our nature and nurture, than that our brains have identical physico-chemical descriptions" (*BS*, p. xvi). So if we adhere to functionalism, a plausible result in cognitive psychology—the discovery that people do not have the same programs—threatens to establish that people do not have beliefs at all.[5]

We can now see one of the principal virtues of Dennett's instrumentalism about intentional systems. Since describing an object as an intentional system entails nothing whatever about either the physico-chemical nature or the functional design of the mechanism that causes the object's behavior, neither neurophysiology nor "sub-personal cognitive psychology" (which studies the functional organization or program of the organism) could possibly show that the object was not an intentional system. Thus if beliefs and

4. Skinner often muddies the waters by claiming to offer "translations" of common-sense mentalistic terms into the language of behaviorism. But, as Dennett and others have noted, (*BS*, pp. 53–70) these "translations" generally utterly fail to capture the meaning or even the extension of the common-sense term being "translated."

5. For an elaboration of the point, cf. Thomas Nagel, "Armstrong on the mind," *Philosophical Review*, 79 (1970), pp. 394–403.

desires (or some respectable *Ersätze*) can be characterized in terms of intentional-system theory, we need have no fear that advances in psychology or brain science might establish that people do not really have beliefs and desires. So the viability of our "conceptual scheme of moral agents" is sustained, in this quarter at least.[6]

Now, finally, it is clear how Dennett's preoccupation with moral themes bears on his eliminative materialism. Recall that Dennett proposes to trade our ungainly folk-psychological notions for concepts characterized in terms of intentional systems. The claim is not that the new concepts are identical with the old, but that they are *better*. They are clearer, more systematic, free from the incoherence lurking in folk notions, *and they capture everything in folk psychology that is worth caring about*. One of the things worth caring about, for Dennett, is the suitability of the clarified notions for incorporation into science. However, if he is to succeed in insulating our moral world-view from the threat posed by scientific psychology, then there is obviously something else Dennett must count as worth caring about. The new concepts built from intentional-system notions must be as serviceable as the older folk notions in sustaining our vision of ourselves as persons.

II

In this section I want to examine just how well Dennett's intentional system *Ersätze* mirror the notions of folk psychology. My focus will be on the "program receptive" notions of belief and desire, concepts which should be easiest to purchase "with intentional coin," and my claim will be that the fit between our common-sense notions and Dennett's proffered replacements is a very poor one.[7] Of course, Dennett does not maintain that the fit is perfect, only that intentional-system theory preserves "the parts of folk psychology worth caring about" (*TK*, p. 30). This is the doctrine I am concerned to challenge. On my view, the move to an intentional-system-characterized notion of belief would leave us unable to say a great deal that we wish to say about ourselves and our fellows. Moreover, the losses will be important ones. If we accept Dennett's trade, we will have no coherent way to describe our cognitive shortcomings nor the process by which we may learn to overcome them. Equally unwelcome, the thriving scientific study of the strengths and weaknesses of human reasoning would wither and die, its hypotheses ruled literally incoherent. What is more, the instrumentalism of Dennett's intentional-system notions seems to fly in the face of some deeply rooted intuitions about responsibility and

6. An entirely parallel strategy works for those other common-sense mental phenomena which Dennett takes to be essential to our concept of ourselves as persons—e.g., consciousness (*BS*, p. 269). If we can give an acceptable intentional system *Ersätz* for the folk-psychological notion of consciousness, we need have no fear that advances in science will threaten our personhood by showing that the notion of consciousness is otiose in the causal explanation of our behavior.

7. For some qualms about Dennett's treatment of "program resistant" features of mentality like pains, see my "Headaches," *Philosophical Books*, April 1980.

moral agency. Throughout most of what follows, I will cleave to the fiction that we already have a tolerably well worked out normative theory of rationality, or could readily build one, though in the closing pages I will offer some skeptical thoughts about how likely this fiction is.

I begin with the problems posed by irrationality. An intentional system, recall, is an ideally rational system; it believes, wants and does just what it ought to, as stipulated by a normative theory of rationality. People, by contrast, are not ideally rational, and therein lies a devastating problem for Dennett. If we were to adopt his suggestion and trade up to the intentional-system notions of belief and desire (hereafter IS belief and IS desire), then we simply would not be able to say all those things we need to say about ourselves and our fellows when we deal with each other's idiosyncrasies, shortcomings, and cognitive growth.

Consider belief. Presumably no system *ought* to hold contradictory beliefs, and all systems *ought* to believe all the logical truths, along with all the logical consequences of what they believe (cf. *BS*, pp. 11, 20, 44; *TK*, p. 11). But people depart from this ideal in a variety of ways. We generally fail to believe *all* logical consequences of our beliefs—sometimes because the reasoning required would be difficult, and sometimes because we simply fail to take account of one or more of our beliefs. Suppose, for example, that an astronaut set the controls incorrectly and has sent his craft into a perilous spin. One possible explanation of his mistake would be that the on-board computer was down, and he had to hand-calculate the setting for the controls. He made a mistake in the calculation, and thus came to have a mistaken belief about what the setting should be. Another possibility is that, although he knew the craft was in the gravitational field of a nearby asteroid—indeed he could see it through the window—he simply forgot to take this into account in figuring out where the control should be set. There is nothing in the least paradoxical about these explanations. We offer similar explanations all the time in explaining our own actions and those of other people. Indeed, since these explanations are so intimately bound up with our notions of excuse and blame, quick-wittedness, absent-mindedness and a host of others, it boggles the mind to try imagining how we would get on with each other if we resolved to renounce them. But if, following Dennett, we agree to swap the folk notion of belief for the intentional-system notion, then renounce them we must. It simply makes no sense to attribute inferential failings or inconsistent beliefs to an ideally rational system.

Our intuitive grasp on the notion of rational desire is rather more tenuous than our grasp on the analogous notion for belief. Still, there seem to be many cases in which we want to ascribe desires to people which are not rational on any plausible reading of that term. Jones is a successful writer, in good health, with many friends and admirers. But he says he wants to die, and ultimately takes his own life. Smith has a dreadful allergy to chocolate, and he knows it. One taste and he is condemned to a week of painful, debilitating hives. But he *really* wants that chocolate bar at the checkout counter. After staring at it for a minute, he buys it and gobbles it down. Brown collects spiders. They are of no

economic value, and he doesn't even think they are very pretty. But it is his hobby. He wants to add to his collection a specimen of a species found only in the desert. So, despite his dislike of hot weather, he arranges to spend his vacation spider hunting in Nevada. By my lights, both Jones's desire and Smith's are simply irrational. As for Brown, "irrational" seems much too strong. Yet it is certainly implausible to say that he *ought* to want that spider. So, on Dennett's account, it is not a rational desire. But idealized intentional systems have all and only the desires they ought to have: Thus if we trade the commonsense notion of want for Dennett's IS want, we simply will not be able to say that Brown wants the spider or that Jones wants to die.

The existence of examples like the ones I have been sketching is not news to Dennett. From his earliest paper on intentional systems to his most recent, he has struggled with analogous cases. Unfortunately, however, he is far from clear on what he proposes to do about them. As I read him, there are two quite different lines that he proposes; I will call them the *hard line* and the *soft line*. Neither is carefully spelled out in Dennett's writings, and he often seems to endorse both within a single paper. Once they have been sharply stated, I think it will be clear that neither line is tenable.

THE HARD LINE

The hard line sticks firmly with the idealized notion of an intentional system and tries to minimize the importance of the gap between IS beliefs and IS desires and their folk-psychological namesakes. The basic ploy here is to suggest that when folk psychology ascribes contradictory beliefs to people or when it insists that a person does not believe some of the consequences of his beliefs, folk psychology undermines its own usefulness and threatens to lapse into incoherence. When this happens, we are forced back to the design stance or the physical stance:

> The presumption of rationality is so strongly entrenched in our inference habits that when our predictions [based on the assumption] prove false, we at first cast about for adjustments in the information possession conditions (he must not have heard, he must not know English, he must not have seen *x*, . . .) or goal weightings, before questioning the rationality of the system as a whole. In extreme cases personalities may prove to be so unpredictable from the intentional stance that we abandon it, and if we have accumulated a lot of evidence in the meanwhile about the nature of response patterns in the individual, we may find that a species of design stance can be effectively adopted. This is the fundamentally different attitude we occasionally adopt toward the insane. (*BS*, pp. 9–10)

Here, surely, Dennett is *just wrong* about what we do when predictions based on idealized rationality prove false. When a neighborhood boy gives me the wrong change from my purchase at his lemonade stand, I do not assume that he believes quarters are only worth

23 cents, nor that he wants to cheat me out of the 2 cents I am due. My *first* assumption is that he is not yet very good at doing sums in his head. Similarly, when a subject working on one of Wason and Johnson-Laird's deceptively difficult reasoning tasks gets the wrong answer, we are not likely to assume that he didn't understand the instructions, nor that he didn't want to get the right answer. Our *first* assumption is that he blew it; he made a mistake in reasoning.[8] What misleads Dennett here is that he is focusing on cases of counter-intuitive or unfamiliar cognitive failings. When someone seems to have made a mistake we can't readily imagine ourselves ever making, we do indeed begin to wonder whether he might perhaps have some unanticipated beliefs and desires. Or if a person seems to be making enormous numbers of mistakes and ending up with a substantial hoard of bizarre beliefs, we grow increasingly reluctant to ascribe beliefs and desires to him at all. Perhaps we count him among the insane. These facts will assume some importance later one. But they are of little use to the hard-line defense of intentional systems. For it is in the diverse domain of more or less familiar inferential shortcomings that common sense most readily and usefully portrays people as departing from an idealized standard of rationality.

Dennett frequently suggests that we cannot coherently describe a person whose beliefs depart from the idealized standard:

> Conflict arises . . . when a person falls short of perfect rationality, and avows beliefs that either are strongly disconfirmed by the available empirical evidence or are self-contradictory or contradict other avowals he has made. If we lean on the myth that a man is perfectly rational, we must find his avowals less than authoritative: "You *can't* mean—understand—what you're saying!"; if we lean on his right as a speaking intentional system to have his word accepted, we grant him an irrational set of beliefs. Neither position provides a stable resting place, for, as we saw earlier, *intentional explanation and prediction cannot be accommodated either to breakdown or to less than optimal design, so there is no coherent intentional description of such an impasse*. (*BS*, 20; last emphasis added)[9]

In the paper from which the quote is taken, Dennett uses 'intentional description,' 'intentional explanation' and the like for both common-sense belief–desire accounts and idealized intentional system accounts. The ambiguity this engenders is crucial in evaluating his claim. On the idealized intentional systems reading it is a tautology that "there is no coherent intentional description of such an impasse." But on the common-sense reading it is simply false. There is nothing at all incoherent about a (common-sense) intentional description of a man who has miscalculated the balance in his checking account!

8. Cf. P. C. Wason and P. N. Johnson-Laird, *The Psychology of Human Reasoning: Structure and Content* (London: Batsford, 1972).

9. For parallel passages, cf. *TB*, p. 19; *R*, p. 74; *BS*, p. 22.

The fact that folk psychology often comfortably and unproblematically views people as departing from the standard of full rationality often looms large in cases where questions of morality and responsibility are salient. Consider the case of Oscar, the engineer. It is his job to review planned operations at the factory and halt those that might lead to explosion. But one day there is an explosion and, bureaucracy being what it is, three years later Oscar is called before a board of inquiry. Why didn't he halt the hazardous operation? It looks bad for Oscar, since an independent expert has testified that the data Oscar had logically entail a certain equation, and it is a commonplace amongst competent safety engineers that the equation is a sure sign of trouble. But Oscar has an impressive defense. Granted the data he had entails the equation, and granted any competent engineer would know that the equation is a sign of trouble. But at the time of the accident neither Oscar nor anyone else knew that the data logically entailed the equation. It was only six months after the accident that Professor Brain at Cambridge proved a fundamental theorem needed to show that the data entail the equation. Without knowledge of the theorem, neither Oscar nor anyone else could be expected to believe that the data entail the equation.

At several places Dennett cites Quine as a fellow defender of the view that the ascription of inconsistent beliefs is problematic.

> To echo a theme I have long cherished in Quine's work, all the evidence—behavioral *and internal*—we acquire for the correctness of one of these ascriptions is not only evidence against the other, but the best sort of evidence. (*R*, p. 74)

However, Dennett misconstrues Quine's point. What Quine urges is not that *any* inconsistency is evidence of bad translation (or bad belief ascription), but rather that *obvious* inconsistency is a sign that something has gone wrong. For Quine, unlike Dennett, sees translation (and belief ascription) as a matter of putting ourselves in our subject's shoes. And the self we put in those shoes, we are too well aware, departs in many ways from the standard of optimal rationality. The point can be made vividly by contrasting Oscar, our safety engineer, with Otto, a lesser functionary. Otto is charged with the responsibility of memorizing a list of contingency plans: if the red light flashes, order the building evacuated; if the warning light goes on, turn the big blue valve; if the buzzer sounds, alert the manager. Now suppose that while he is on duty the red light flashes but Otto fails to order an evacuation. There is a strong prima-facie case that Otto is to be held responsible for the consequences. Either he failed to see the light (he was asleep or not paying due attention), or he did not memorize the contingency plans as he was obligated to, or he has some sinister motive. But, and this is the crucial point, it will be no excuse for Otto to claim that he had memorized the plan, saw the light, and was paying attention, but it just never occurred to him to order the evacuation. It is in these cases of apparently blatant or "incomprehensible" irrationality that we hunt first for hidden motives or beliefs. For, absent these, the subject must be judged irrational in a way we cannot imagine

ourselves being irrational; and it is this sort of irrationality that threatens the application of our common-sense notions of belief and desire.

In Dennett's writings there are frequent hints of a second strategy for defending the hard line, a strategy which relies on an evolutionary argument. He cheerfully concedes that he has "left [his] claim about the relation between rationality and evolutionary considerations so open-ended that it is hard to argue against efficiently" (*R*, p. 73). Still, I think it is important to try wringing some arguments out of Dennett's vague meditations on this topic. As I read him, Dennett is exploring a pair of ideas for showing that the gap between IS notions and folk notions is much smaller than some have feared. If he can show this, the hard line will have been vindicated.

The first idea is suggested by a passage (*BS*, pp. 8–9) in which Dennett asks whether we could adopt the intentional stance toward exotic creatures encountered on an alien planet. His answer is that we could, provided "we have reason to suppose that a process of natural selection has been in effect . . ." (*BS*, p. 8). The argument seems to be that natural selection favors true beliefs, and thus will favor cognitive processes which generally yield true beliefs in the organism's natural environment. So if an organism is the product of natural selection, we can safely assume that most of its beliefs will be true, and most of its belief-forming strategies will be rational. Departures from the normative standard required by the intentional stance will be few and far between.

For two quite different reasons, this argument is untenable. First, it is simply not the case that natural selection favors true beliefs over false ones. What natural selection does favor is beliefs which yield selective advantage. And there are many environmental circumstances in which false beliefs will be more useful than true ones. In these circumstances, natural selection ought to favor cognitive processes which yield suitable false beliefs and disfavor processes which yield true beliefs. Moreover, even when having true beliefs is optimal, natural selection may often favor a process that yields false beliefs most of the time, but which has a high probability of yielding true beliefs when it counts. Thus, for example, in an environment with a wide variety of suitable foods, an organism may do very well if it radically overgeneralizes about what is inedible. If eating a certain food caused illness on a single occasion, the organism would immediately come to believe (falsely, let us assume) that all passingly similar foods are poisonous as well. When it comes to food poisoning, *better safe than sorry* is a policy that recommends itself to natural selection.[10]

The second fault in the argument I am attributing to Dennett is a subtle but enormously important one. As stated, the argument slips almost unnoticeably from the claim that natural selection favors cognitive processes which yield true beliefs in the natural environment to the claim that natural selection favors *rational* belief-forming strategies.

10. For a detailed discussion of some examples and further references, cf. H. A. Lewis, "The Argument From Evolution," *Proceedings of the Aristotelian Society*, Supplementary vol. LIII, 1979; also my "Could man be an irrational animal?" *Synthese* 64 (1985), 115–35.

But, even if the first claim were true, the second would not follow. There are many circumstances in which inferential strategies which from a normative standpoint are patently invalid will nonetheless generally yield the right answer. The social-psychology literature is rich with illustrations of inferential strategies which stand subjects in good stead ordinarily but which subjects readily overextend, with unhappy results.[11]

So long as we recognize a distinction between a normative theory of inference or decision making and a set of inferential practices which (in the right environment) generally get the right (or selectively useful) answer, it will be clear that the two need not, and generally do not, coincide. However, in a number of places Dennett seems to be suggesting that there really *is* no distinction here, that by "normative theory of inference and decision" he simply *means* "practices favored by natural selection." This move is at the core of the second idea I see in Dennett for using evolutionary notions to buttress the hard line (Cf. *R*, pp. 73–4). And buttress it would! For it would then become *tautologous* that naturally evolved creatures are intentional systems, believing, wanting and doing what they ought, save when they are malfunctioning. Yet Dennett will have to pay a heavy price for turning the hard line into a tautology. For if *this* is what he means by "normative theory of belief and decision," then such established theories as deductive and inductive logic, decision theory and game theory are of no help in assessing what an organism "ought to believe." Natural selection, as we have already noted, sometimes smiles upon cognitive processes that depart substantially from the canons of logic and decision theory. So these established theories and our guesses about how to extend them will be of no help in assessing what an intentional system should believe, desire or do. Instead, to predict from the intentional stance we should need a detailed study of the organism's physiology, its ecological environment and its history. But predicting from the intentional stance, characterized in *this* way, is surely not to be recommended when we "doubt the practicality of prediction from the design or physical stance" (*BS*, p. 8). Nor, obviously, does *this* intentional stance promise to yield belief and desire attributions that are all but co-extensive with those made in common sense.

This is all I shall have to say by way of meeting the hard line head on, I think it is fair to conclude that the hard line simply cannot be maintained. The differences separating the IS notions of belief and desire from their common-sense counterparts is anything but insubstantial. Before turning to Dennett's soft line, we should note a further unwelcome consequence of rejecting folk psychology in favor of intentional-system theory. During the last decade, cognitive psychologists have become increasingly interested in studying the strengths and foibles of human reasoning. There is a substantial and growing literature aimed at uncovering predictable departures from normative standards of reasoning and decision making, almost all of it implicitly or explicitly cast in the idiom of folk

11. Cf. Richard Nisbett and Lee Ross, *Human Inference* (Englewood Cliffs, NJ: Prentice-Hall, 1980).

psychology.[12] Were we to replace folk notions with their intentional-system analogs, we should have to conclude that all of this work limning the boundaries of human rationality is simply incoherent. For, as Dennett notes, "the presuppositions of intentional explanation . . . put prediction of *lapses* in principle beyond its scope . . ." (*BS*, p. 246).[13]

THE SOFT LINE

In contrast with the hard line, which tries to minimize the size or importance of the difference between folk and IS notions, the soft line acknowledges a substantial and significant divergence. To deal with the problems this gap creates, the soft line proposes some fiddling with the idealized notion of an intentional system. The basic idea is that once we have an idealized theory of intentional systems in hand, we can study an array of variations on the idealized theme. We can construct theories about "imperfect intentional systems" (the term is mine, not Dennett's) which have specified deficiences in memory, reasoning power, etc. And we can attempt to determine empirically which imperfect intentional system best predicts the behavior of a particular subject or species. Rather than assuming the intentional stance toward an organism or person, we may assume one of a range of "imperfect intentional stances," from which it will make sense to ascribe a less than fully rational set of beliefs and desires. From these various stances we can give intentional descriptions of our cognitive shortcomings and elaborate an empirical science which maps the inferential strengths and weaknesses of humans and other creatures. We can also legitimize our folk-psychological descriptions of ourselves—protecting "personhood from the march of science" (*R*, p. 75)—by appeal to the imperfect-intentional-system theory which best predicts our actual behavior. But *genuine* intentional-system theory (*sans phrase*) would have a definite pride of place among these theories of imperfect intentional systems. For all of the latter would be variations on the basic IS framework.

Dennett, with his disconcerting penchant for working both sides of the street, never flatly endorses the soft line, though it is clear that he has pondered something like it:

> Consider a set T of transformations that take beliefs into beliefs. The problem is to determine the set T_s for each intentional system S, so that if we know that S believes

12. E.g., Nisbett and Ross, ibid., and Wason and Johnson-Laird, ibid., along with the many studies cited in these books.

13. Dennett appends the following footnote to the quoted sentence: "In practice we predict lapses at the intentional level ('You watch! He'll forget all about your knight after you move the queen') on the basis of loose-jointed inductive hypotheses about individual or widespread human frailties. These hypotheses are expressed in intentional terms, but if they were given rigorous support, they would in the process be recast as predictions from the design or physical stance" (*BS*, p. 246). So the scientific study of intentionally described inferential shortcomings can aspire to no more than "loose-jointed hypotheses" in need of recasting. But cf. *TK*, pp. 11–12, where Dennett pulls in his horns a bit.

> *p*, we will be able to determine other things that *S* believes by seeing what the transformations of *p* are for T_s. If *S* were ideally rational, every valid transformation would be in T_s; *S* would believe every logical consequence of every belief (and, ideally, *S* would have no false beliefs). Now we know that no actual intentional system will be ideally rational; so we must suppose any actual system will have a *T* with less in it. But we also know that, to qualify as an intentional system at all, *S* must have a *T* with some integrity; *T* cannot be empty. (*BS*, p. 21)

In the next few sentences, however, Dennett expresses qualms about the soft line:

> What rationale could we have, however, for fixing some set between the extremes and calling it *the* set for belief (for *S*, for earthlings, for ten-year-old girls)? This is another way of asking whether we could replace Hintikka's normative theory of belief with an empirical theory of belief, and, if so, what evidence we would use. "Actually," one is tempted to say, "people do believe contradictions on occasion, as their utterances demonstrate; so any adequate logic of belief or analysis of the concept of belief must accommodate this fact." But any attempt to *legitimize* human fallibility in a theory of belief by fixing a permissible level of error would be like adding one more rule to chess: an Official Tolerance Rule to the effect any game of chess containing no more than *k* moves that are illegal relative to the other rules of the game is a legal game of chess. (*BS*, p. 21)

In a more recent paper, Dennett sounds more enthusiastic about the soft line:

> *Of course* we don't all sit in the dark in our studies like mad Leibnizians rationalistically excogitating behavioral predictions from pure, idealized concepts of our neighbors, nor do we derive all our readiness to attribute desires to careful generation of them from the ultimate goal of survival. . . . Rationalistic generation of attributions is augmented and even corrected on occasion by empirical generalizations about belief and desire that guide our attributions and are learned more or less inductively. . . . I grant the existence of all this naturalistic generalization, and its role in the normal calculation of folk psychologist—i.e., all of us. . . . *I would insist, however, that all this empirically obtained lore is laid over a fundamental generative and normative framework that has features I have described*. (*TK*, pp. 14–15, last emphasis added)

Whatever Dennett's considered view may be, I think the soft line is clearly preferable to the hard line. Indeed, the soft line is similar to a view that I have myself defended.[14] As

14. In "On the ascription of content," in A. Woodfield (ed.), *Thought and Object* (Oxford University Press, 1982).

a way of focusing in on my misgivings about the soft line, let me quickly sketch my own view and note how it differs from the view I am trying to foist on Dennett. Mine is an effort squarely situated in what Dennett calls "the anthropological quest" (*TK*, p. 6). I want to describe as accurately as possible just what we are up to when we engage in the "folk practice" of ascribing beliefs to one another and dealing with one another partly on the basis of these ascriptions. My theory is an elaboration on Quine's observation that in ascribing beliefs to others "we project ourselves into what, from his remarks and other indications, we imagine the speaker's state of mind to have been, and then we say what, in our language, is natural and relevant for us in the state thus feigned" (*World and Object*, p. 219). As I see it, when we say *S believes that p* we are saying that *S* is in a certain sort of functionally characterized psychological state, viz., a "belief state." The role of the "content sentence," *p*, is to specify *which* belief state it is. If we imagine that we ourselves were not to utter *p* in earnest, the belief we are attributing to *S* is one *similar* (along specified dimensions) to the belief which would cause our own imagined assertion. One of the dimensions of similarity that figures in belief ascription is the pattern of inference that the belief states in question enter into. When the network of potential inferences surrounding a subject's belief state differs substantially from the network surrounding our own belief that *p*, we are reluctant to count the subject's belief as a belief *that p*. Thus we will not have any comfortable way of ascribing content to the belief states of a subject whose inferential network is markedly different from ours. Since we take ourselves to approximate rationality, this explains the fact, noted by Dennett, that intentional description falters in the face of egregious irrationality. It also explains the fact, missed by Dennett, that familiar irrationality—the sort we know ourselves to be guilty of—poses no problem for folk psychology.

A full elaboration of my theory would be a long story, out of place here. What is important for our present purposes is to note the differences between my account and what I have been calling Dennett's soft line. These differences are two. First, my story does not portray folk psychology as an *instrumentalist* theory. Belief states are *functional* states which can and do play a role in the causation of behavior. Thus folk psychology is not immune from the advance of science. If it turns out that the human brain does not have the sort of functional organization assumed in our folk theory, then there are no such things as beliefs and desires. Second, the notion of idealized rationality plays *no role at all* in my account. In ascribing content to belief states, we measure others not against an idealized standard but against ourselves. It is in virtue of this Protagorean parochialism that the exotic and the insane fall outside the reach of intentional explanation.

So much for the difference between my view and Dennett's. Why should mine be preferred? There are two answers. First, I think it is simply wrong that we ordinarily conceive of beliefs and desires in instrumentalist terms—as abstracta rather than illata. It is, however, no easy task to take aim at Dennett's instrumentalism, since the target refuses to stay still. Consider:

> Folk psychology is *instrumentalistic* . . . Beliefs and desires of folk psychology . . . are abstracta. (*TK*, p. 13)

> It is not particularly to the point to argue against me that folk psychology is *in fact* committed to beliefs and desires as distinguishable, causally interacting *illata*; what must be shown is that it ought to be. The latter claim I will deal with in due course. The former claim I *could* concede without embarrassment to my overall project, but I do not concede it, for it seems to me that the evidence is quite strong that our ordinary notion of belief has next to nothing of the concrete in it. (*TK*, p. 15)

> The *ordinary* notion of belief no doubt does place beliefs somewhere midway between being *illata* and being *abstracta*. (*TK, p*. 16)

In arguing for his sometimes instrumentalism Dennett conjures the sad tale of Pierre, shot dead by Jacques in Trafalgar Square. Jacques

> is apprehended on the spot by Sherlock; Tom reads about it in the *Times* and Boris learns of it in *Pravda*. Now Jacques, Sherlock, Tom and Boris have had remarkably different experiences—to say nothing of their earlier biographies and future prospects—but there is one thing they share: they all believe that a Frenchman has committed a murder in Trafalgar Square. They did not all *say* this, not even "to themselves"; *that proposition* did not, we can suppose, "occur to" any of them, and even if it had, it would have had entirely different import for Jacques, Sherlock, Tom and Boris. (*TK*, p. 15)

Dennett's point is that while all four men believe that a Frenchman committed a murder in Trafalgar Square, their histories, interests and relations to the deed are so different that they could hardly be thought to share a single, functionally characterizable state. This is quite right, but it does not force us to view beliefs as abstracta. For if, as my theory insists, there is a *similarity* claim embedded in belief ascriptions, then we should expect these ascriptions to be both vague and sensitive to pragmatic context. For Jacques and Boris both to believe that a Frenchman committed a murder in Trafalgar Square, they need not be in the very same functional state, but only in states that are sufficiently similar for the communicative purposes at hand.

As Dennett notes, one need not be crucially concerned with what "folk psychology is in fact committed to." Since he aims to replace folk psychology with intentional-system notions, it would suffice to show that the instrumentalism of these latter notions is no disadvantage. But here again I am skeptical. It is my hunch that our concept of ourselves as moral agents simply will not sit comfortably with the view that beliefs and desires are mere computational conveniences that correspond in no interesting way to what goes on inside the head. I cannot offer much of an argument for my hunch,

though I am encouraged by the fact that Dennett seems to share the intuition lying behind it:

> Stich accurately diagnoses and describes the strategic role I envisage for the concept of an intentional system, permitting the claim that human beings are genuine believers and desirers to survive almost any imaginable discoveries in cognitive and physiological psychology, thus making our status as moral agents well nigh invulnerable to scientific discontinuation. Not 'in principle' invulnerable, for in a science-fiction on mood we can imagine startling discoveries (e.g., some 'people' are organic puppets remotely controlled by Martians) that would upset any particular home truths about believers and moral agenthood you like. . . . (*R*, p. 73)

Now if our concept of moral agenthood were really compatible with the intentional-system construal of beliefs and desires, it is hard to see why the imagined discovery about Martians should be in the least unsettling. For, controlled by Martians or not, organic puppets are still intentional systems in perfectly good standing. So long as their behavior is usefully predictable from the intentional stance, the transceivers inside their heads sanction no skepticism about whether they really have IS beliefs and IS desires. But Dennett is right, of course. We would not count his organic puppets as believers or moral agents. The reason, I submit, is that the morally relevant concept of belief is not an instrumentalistic concept.

The second reason for preferring my line to Dennett's soft line is that the idea of a *normative* theory of beliefs and desires, which is central to Dennett's view, plays no role in mine. And this notion, I would urge, is one we are best rid of. Recall that from the outset we have been relying on rough and ready intuitions about what an organism ought to believe, desire and do, and assuming that these intuitions could be elaborated and systematized into a theory. But I am inclined to think that this assumption is mistaken. Rather, it would appear that the intuitions Dennett exploits are underlain by a variety of different ideas about what an organism ought to believe or desire, ideas which as often as not pull in quite different directions. Sometimes it is an evolutionary story which motivates the intuition that a belief or desire is the one a well-designed intentional system should have. At other times intuitions are guided by appeal to logic or decision theory. But as we have seen, the evolutionary account of what an organism ought to believe and desire just will not do for Dennett, since it presupposes an abundance of information about the ecological niche and physiological workings of the organism. Nor is there any serious prospect of elaborating logic and decision theory into a suitably general account of what an organism ought to believe and desire. Indeed, apart from a few special cases, I think our intuitions about what an organism ought to believe and desire are simply nonexistent. The problem is not merely that we lack a worked-out normative theory of belief and desire; it runs much deeper. For in general we have no idea what such a normative theory would be telling us. We do not really know what it

means to say that an organism *ought to have* a given belief or desire. Consider some examples:

Ought Descartes to have believed his theory of vortices?
Ought Nixon to have believed that he would not be impeached?
Ought William James to have believed in the existence of a personal God?
Should all people have perfect memories, retaining for life all beliefs save those for which they later acquire negative evidence?

In each of these cases our grasp of what the question is supposed to *mean* is at best tenuous. The prospects of a *general theory* capable of answering all of them in a motivated way are surely very dim. Worse still, the general theory of intentional systems that Dennett would have us work toward must tell us not only what *people* in various situations ought to believe, but also what other animals ought to believe. Ought the frog to believe that there is an insect flying off to the right? Or merely that there is some food there? Or perhaps should it only have a conditional belief: if it flicks its tongue in a certain way, something yummy will end up in its mouth? Suppose the fly is of a species that causes frogs acute indigestion. Ought the frog to believe this? Does it make a difference how many fellow frogs he has seen come to grief after munching on similar bugs? A normative theory of desire is, if anything, more problematic. Should I want to father as many offspring as possible? Should the frog?

To the extent that these questions are obscure, the notion of a normative theory of belief and desire is obscure. And that obscurity in turn infects much of what Dennett says about intentional systems and the intentional stance. Perhaps Dennett can dispel some of the mystery. But in the interim I am inclined to think that the normatively appropriate attitude is the skepticism I urged in my opening paragraph.[15]

15. I have learned a good deal from the helpful comments of Bo Dahlbom, Robert Cummins, Philip Pettit and Robert Richardson.