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.

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 *Ersatz* 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.

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.

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.

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

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.

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

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

Making Sense of Ourselves

DANIEL C. DENNETT

Stich has (again[1]) given a lively, sympathetic, and generally accurate account of my view and once again he disagrees, this time with more detailed objections and counterproposals. My proposed refinement of the folk notion of belief (via the concept of an *intentional system*) would, he claims, "leave us unable to say a great deal that we now wish to say about ourselves." For this to be an objection, he

"Making Sense of Ourselves" by D. C. Dennett first appeared in *Philosophical Topics* 12 (1981), pp. 63–81, and is reprinted here by permission.

must mean it would leave us unable to say a great deal we *rightly* want to say – because it is true, presumably. We must see what truths, then, he supposes are placed out of reach by my account. Many of them lie, he says, in the realm of facts about our cognitive shortcomings, which can be given no coherent description according to my account: "if we trade up to the intentional-system notions of belief and 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" (this volume, p. 173). He gives several examples. Among them are the forgetful astronaut, the boy at the lemonade stand who gives the wrong change, and the man who has miscalculated the balance in his checking account. These three are cases of simple, unmysterious cognitive failure – cases of people *making mistakes* – and Stich claims that my view cannot accommodate them. One thing that is striking about all three cases is that in spite of Stich's summary expression of his objection, these are *not* cases of "familiar irrationality" or cases of "inferential failings" at all. They are not cases of what we would *ordinarily* call irrationality, and since there are quite compelling cases of what we *would* ordinarily call irrationality (and since Stich knows them and indeed cites some of the best documented cases[2]), it is worth asking why he cites instead these cases of miscalculation as proof against my view. I shall address this question shortly, but first I should grant that these are in any case examples of suboptimal behavior of the sort my view is not supposed to be able to handle.

I hold that such errors, as either *malfunctions* or the outcomes of *misdesign*, are unpredictable from the intentional stance, a claim with which Stich might agree, but I go on to claim that there will inevitably be an instability or problematic point in the mere *description* of such lapses at the intentional system level – at the level at which it is the agent's beliefs and desires that are attributed. And here it seems at first that Stich must be right. For although we seldom if ever suppose we can *predict* people's particular mistakes from our ordinary folk-psychological perspective, there seems to be nothing more straightforward than the folk-psychological *description* of such familiar cases. This presumably is part of the reason why Stich chose these cases: they are so uncontroversial.

Let's look more closely, though, at one of the cases, adding more detail. The boy's sign says "LEMONADE – 12 cents a glass." I hand him a quarter, he gives me a glass of lemonade and then a dime and a penny change. He's made a mistake. Now what can we *expect* from him when we point out his error to him? That he will exhibit surprise, blush, smite his forehead, apologize, and give me two cents. Why do we expect him to exhibit surprise? Because we attribute to him the belief that he's given me the right change – he'll be surprised to learn that he hasn't.[3] Why do we expect him to blush? Because we attribute to him the desire not to cheat (or be seen to cheat) his customers. Why do we expect him to smite his forehead or give some other acknowledgment of his lapse? Because we attribute to him not only the belief that 25 - 12 = 13, but also the belief that that's obvious, and the belief that no one his age should make any mistakes about it. While we can't predict his particular error – though we might have made an actuarial prediction that he'd probably make some such error before the day was out – we can pick up the skein of our intentional interpretation once he has made his mistake and predict his further reactions and activities with no more than the

usual attendant risk. At first glance then it seems that belief attribution in this instance is as easy, predictive and stable as it ever is.

But now look yet more closely. The boy has made a mistake all right, but *exactly which mistake*? This all depends, of course, on how we tell the tale – there are many different possibilities. But no matter which story we tell, we will uncover a problem. For instance, we might plausibly suppose that so far as all our evidence to date goes, the boy believes:

1 that he has given me the right change
2 that I gave him a quarter
3 that his lemonade costs 12 cents
4 that a quarter is 25 cents
5 that a dime is 10 cents
6 that a penny is 1 cent
7 that he gave me a dime and a penny change
8 that 25 - 12 = 13
9 that 10 + 1 = 11
10 that 11 ≠ 13

Only (1) is a false belief, but how can he be said to believe *that* if he believes all the others? It surely is not plausible to claim that he has *mis-inferred* (1) from any of the others, directly or indirectly. That is, we would not be inclined to attribute to him the inference of (1) directly from (7) and – what? Perhaps he would infer

11 that he gave me 11 cents change

from (9) and (7) – he *ought to*, after all – but *it would not make sense* to suppose he *inferred* (1) from (11) unless he were under the misapprehension

12 that 11 cents is the right change from a quarter.

We would expect him to believe *that* if he believed

13 that 25 - 12 = 11

and while we *might* have told the tale so that the boy simply had this false belief – and *didn't* believe (8) – (we can imagine, for instance, that he thought that's what his father told him when he asked), this would yield us a case that was not at all a plausible case of either irrationality or even miscalculation, but just a case of a perfectly rational thinker with a single false belief (which then generates other false beliefs such as (1)). Stich rightly does not want to consider such a case, for of course I do acknowledge the possibility of mere false belief, when special stories can be told about its acquisition. If we then attribute (13) *while retaining* (8) we get a blatant and bizarre case of irrationality: someone believing simultaneously that 25 – 12 = 13, 25 - 12 = 11 and 13 ≠ 11. This is not what we had supposed

at all, but so strange that we are bound to find the conjoined attributions frankly incredible. Something has to give. If we say, as Stich proposes, that the boy "is not yet very good at doing sums in his head" what is the implication? That he doesn't *really* believe the inconsistent triad, that he *sort* of understands arithmetical notions well enough to have the cited beliefs? That is, if we say what Stich says and *also* attribute the inconsistent beliefs, we still have the problem of brute irrationality too stark to countenance; if we take Stich's observation to temper or withdraw the attribution, then Stich is agreeing with me: even the simplest and most familiar errors require us to resort to scare-quotes or other *caveats* about the literal truth of the total set of attributions.

There is something obtuse, of course, about the quest exhibited above for a total belief-set surrounding the error. The demand that we find an inference – even a *mis*-inference – to the false belief (1) is the demand that we find a practice or tendency with something like a rationale, an exercise of which has led in this instance to (1). No mere succession in time or even regular causation is enough in itself to count as an inference. For instance, were we to learn that the boy was led directly from his belief (6) that a penny is 1 cent to his belief (2) that I gave him a quarter, then no matter how habitual and ineluctable the passage in him from (6) to (2), we wouldn't call it *inference*.[4] Inferences are passages of thought for which there is a reason, but people don't make mistakes for reasons. Demanding reasons (as opposed to "mere" causes) for mistakes generates spurious edifices of belief, as we have just seen in (11–13), but simply acquiescing in the attribution of reasonless belief is no better. It is not as if *nothing* led the boy to believe (1); it is not as if that belief was utterly baseless. We do not suppose, for instance, that he would have believed (1) had his hand been empty, or filled with quarters, or had I given him a dollar or a credit card. He does somehow base his mistaken belief on a distorted or confused or mistaken perception of what he is handing me, what I have handed him, and the appropriate relationships between them.

The boy is basically on top of the situation, and is no mere change-giving robot; nevertheless, we must descend from the level of beliefs and desires to some other level of theory to describe his mistake, since no account in terms of his beliefs and desires will make sense completely. At some point our account will have to cope with the sheer senselessness of the transition in any error.

My perhaps tendentious examination of a single example hardly constitutes an argument for my general claim that this will always be the outcome. It is presented as a challenge: try for yourself to tell the total belief story that surrounds such a simple error and see if you do not discover just the quandary I have illustrated.

Mistakes of the sort exhibited in this example are slips in good procedures, not manifestations of an allegiance to a bad procedure or principle. The partial confirmation of our inescapable working hypothesis that the boy is fundamentally rational is his blushing acknowledgement of his error. He doesn't defend his action once it is brought to his attention, but willingly corrects his error. This is in striking contrast to the behavior of agents in the putative cases of genuine irrationality cited by Stich. In these instances, people not only persist in their "errors," but stubbornly defend their practice – and find defenders among philosophers as well.[5] It is at least *not obvious* that there are any cases of systematically irrational behavior or thinking. The cases that have been proposed

are all controversial, which is just what my view predicts; no such thing as a cut-and-dried or obvious case of "familiar irrationality." This is not to say that we are always rational, but that when we are not, the cases defy description in ordinary terms of belief and desire. There is no mystery about why this should be so. An intentional interpretation of an agent is an exercise that attempts to *make sense* of the agent's acts, and when acts occur that make no sense, they cannot be straightforwardly interpreted in sense-making terms. Something must give: we allow that the agent either only "sort of" believes this or that, or believes this or that "for all practical purposes," or believes some falsehood which creates a context in which what had appeared to be irrational turns out to be rational after all. (See, e.g., Cohen's suggestions, op. cit, n. 5). These particular fall-back positions are themselves subject to the usual tests on belief attribution, so merely finding a fall-back position is not confirming it. If it is disconfirmed, the search goes on for another saving interpretation. If there is no *saving* interpretation – if the person in question is irrational – no interpretation at all will be settled on.

The same retreat from the abyss is found in the simple cases of miscalculation and error of which Stich reminds us, but with a few added wrinkles worth noting. In the case of the lemonade seller, we might excuse ourselves from further attempts to sort out his beliefs by just granting that while he knew (and thus believed)[6] all the right facts, he "just forgot" or "overlooked" a few of them temporarily – until we reminded him of them. This has the appearance of being a modest little psychological hypothesis: something roughly to the effect that although something or other was stored safe and sound inside the agent's head where it belonged, its address was temporarily misplaced. Some such story may well in the end be supported within a confirmed and detailed psychological theory,[7] but it is important to note that at the present time we make these hypotheses *simply* on the basis of our abhorrence of the vacuum of contradiction.

For instance, consider absentmindedness – a well-named affliction, it seems. At breakfast I am reminded that I am playing tennis with Paul instead of having lunch today. At 12:45 I find myself polishing off dessert when Paul, in tennis gear, appears at my side and jolts me into recollection. "It completely slipped my mind!" I aver, blushing at my own absentmindedness. But why do I say *that*? Is it because, as I recall, not a single conscious thought about my tennis date passed through my head after breakfast? That might be true, but perhaps no conscious thought that I was going to lunch today occurred to me in the interim either, and yet here I am, finishing my lunch. Perhaps if I *had* thought consciously about going to lunch as usual, that very thought would have reminded me that I wasn't, in fact. And in any case, even if I remember now that it *did* once occur to me in mid-morning that I was to play tennis today – to no avail, evidently – I will still say it subsequently slipped my mind.

Why, indeed, am I eager to *insist* that it completely slipped my mind? To assure Paul that I haven't stood him up on purpose? Perhaps, but that should be obvious enough not to need saying, and if my eagerness is a matter of not wanting to insult him, I am not entirely succeeding, since it is not at all flattering to be so utterly forgotten. I think a primary motive for my assertion is just to banish the possibility that otherwise would arise: I am starkly irrational; I believe both that I am playing tennis at lunch and that I am free to go to lunch as usual. I cannot act on both beliefs at once; whichever I act on, I declare the other to have slipped my mind. Not on any introspective evidence (for I may, after all, have *repeatedly* thought of the matter in the relevant interim period), but on *general principles*. It does not matter how close to noon I have reflected on my tennis date; if I end up having lunch as usual the tennis date *must have* slipped my mind at the last minute.

There is no direct relationship between one's conscious thoughts and the occasions when we will say something has slipped one's mind. Suppose someone asks me to have lunch today and I reply that I can't: I have another appointment then, but for the life of me I can't recall what it is – it will come to me later. Here although in one regard my tennis date has slipped my mind, in another it has not, since my belief that I am playing tennis, while not (momentarily) consciously retrievable, is yet doing some work for me: it is keeping me from making the conflicting appointment. I hop in my car and I get to the intersection: left takes me home for lunch; right takes me to the tennis court; I turn right this time without benefit of an accompanying conscious thought to the effect that I am playing tennis today at lunchtime. It has not slipped my mind, though; had it slipped my mind, I would no doubt have turned left.[8] It is even possible to have something slip one's mind while one is thinking of it consciously! "Be careful of this pan," I say, "it is very hot" – reaching out and burning myself on the very pan I am warning about. The height of absentmindedness, no doubt, but possible. We would no doubt say something like "You didn't think what you were saying!" – which doesn't mean that the words issued from my mouth as from a zombie, but that if I had believed – *really* believed – what I was saying, I *couldn't* have done what I did. If I can in this manner not think what I am saying, I could also in about as rare a case not think what I was thinking. I could think "careful of that hot pan" *to myself*, while ignoring the advice.

There is some temptation to say that in such a case, while I knew full well that the pan was hot, I just forgot for a moment. Perhaps we want to acknowledge this sort of forgetting, but note that it is not at all the forgetting we suppose to occur when we say I have forgotten the telephone number of the taxicab company I called two weeks ago, or forgotten the date of Hume's birth. In those cases we presume the information is gone for good. Reminders and hints won't help me recall. When I say "I completely forgot our tennis date," I don't at all mean I completely forgot it – as would be evidenced if on Paul's arrival in tennis gear I was blankly baffled by his presence, denying any recollection of having made the date.

Some other familiar locutions of folk psychology are in the same family: 'notice', 'overlook', 'ignore', and even 'conclude.' One's initial impression is that these terms are applied by us to our own cases on the basis of direct introspection. That is, we classify various conscious acts of our own as concludings, noticings, and the like – but what about ignorings and overlookings? Do we find ourselves doing these things? Only retrospectively, and in a self-justificatory or self-critical mood: "I ignored the development of the pawns on the queen side" says the chess player, "because it was so clear that the important development involved the knights on the king side." Had he lost the game, he would have said "I simply overlooked the development of the pawns on the queen side, since I was under the misapprehension that the king side attack was my only problem."

Suppose someone asks, "Did you *notice* the way Joe was evading your questions

yesterday?" I might answer, "yes," even though I certainly did not *think any conscious thoughts* at the time (that I can recall) about the way Joe was evading my questions; if I can nevertheless see that my reactions to him (as I recall them) took appropriate account of his evasiveness, I will (justly) aver that I did notice. Since I did the appropriate thing in the circumstances, I must have noticed, mustn't I?

In order just now for you to get the gist of my tale of absentmindedness, you had to conclude from my remark about "polishing off dessert" that I had just finished a lunch and missed my tennis date. And surely you did so conclude, but did you *consciously* conclude? Did anything remotely like "Hmm, he must have had lunch . . ." run through your head? Probably not. It is no more likely that the boy selling lemonade consciously thought that the eleven cents in his hand was the right change. "Well, if he didn't *consciously* think it, he unconsciously thought it; we must posit an unconscious controlling thought to that effect to explain, or ground, or *be* (!) his belief that he is giving the right change."

It is tempting to suppose that when we retreat from the abyss of irrationality and find a different level of explanation on which to flesh out our description of errors (or, for that matter, of entirely felicitous passages of thought), the arena we properly arrive at is the folk-psychological arena of thinkings, concludings, forgettings, and the like – not mere abstract mental *states* like belief, but concrete and clockable episodes or activities or processes that can be modeled by psychological model-builders and measured and tested quite directly in experiments. But as the examples just discussed *suggest* (though they do not by any means *prove*), we would be unwise to model our serious, academic psychology too closely on these putative illata of folk theory. We postulate all these apparent activities and mental processes *in order to make sense* of the behavior we observe – in order, in fact, to make as much sense as possible of the behavior, especially when the behavior we observe is our own. Philosophers of mind used to go out of their way to insist that one's access to one's own case in such matters is quite unlike one's access to others', but as we learn more about various forms of psycho-pathology and even the foibles of apparently normal people,[9] it becomes more plausible to suppose that although there are still some small corners of unchallenged privilege, some matters about which our authority is invincible, each of us is in most regards a sort of inveterate auto-psychologist, effortlessly *inventing* intentional interpretations of our own actions in an inseparable mix of confabulation, retrospective self-justification and (on occasion, no doubt) good theorizing. The striking cases of confabulation by subjects under hypnosis or suffering from various well-documented brain disorders (Korsakoff's syndrome, split brains, various "agnosias") raise the prospect that such virtuoso displays of utterly unsupported self-interpretation are not manifestations of a skill suddenly learned in response to trauma, but of a normal way of life unmasked.[10]

As creatures of our own attempts to make sense of ourselves, the putative mental activities of folk theory are hardly a neutral field of events and processes to which we can resort for explanations when the normative demands of intentional system theory run afoul of a bit of irrationality. Nor can we suppose their counterparts in a developed cognitive psychology, or even their "realizations" in wetware of the brain, will fare better. Stich holds out the vision of an entirely norm-free, naturalized psychology that can *settle* the indeterminacies of intentional

system theory by appeal, ultimately, to the presence or absence of real, functionally salient, causally potent states and events that can be identified and *ascribed content independently of the problematic canons of ideal rationality my view requires*. What did the lemonade seller *really believe*? Or what, in any event, was the *exact content* of the sequence of states and events that figure in the cognitivistic description of his error? Stich supposes we will be able, in principle, to say, even in cases where my method comes up empty-handed. I claim, on the contrary, that just as the interpretation of a bit of *outer*, public communication – a spoken or written utterance in natural language, for instance – *depends on* the interpretation of the utterer's beliefs and desires, so the interpretation of a bit of *inner*, sub-personal cognitivistic machinery must inevitably depend on exactly the same thing: the whole person's beliefs and desires. Stich's method of content ascription depends on mine, and is not an alternative, independent method.

Suppose we find a mechanism in Jones that reliably produces an utterance of 'It is raining' whenever Jones is queried on the topic and it is raining in Jones' epistemically accessible vicinity. It also produces 'yes' in response to 'Is it raining?' on those occasions. Have we discovered Jones' belief that it is raining? That is, more circumspectly, have we found the mechanism that "subserves" this belief in Jones' cognitive apparatus? Maybe – it all depends on whether or not Jones believes that it is raining when (and only when) this mechanism is "on." That is, perhaps we have discovered a weird and senseless mechanism (like the "assent-inducing tumor" I imagined in "Brain writing and mind reading," *Brainstorms*, p. 44) that deserves no intentional interpretation at all – or at any rate not this one: that it is the belief that it is raining. We need a standard against which to judge our intentionalistic labels for the illata of sub-personal cognitive theory; what we must use for this standard is the system of abstracta that fixes belief and desire by a sort of hermeneutical process that tells the best, most rational, story that can be told. If we find that Jones passes the right tests – he demonstrates that he really understands what the supposition that it is raining means, for instance – we may find confirmation of our hypothesis that we have uncovered the mechanistic realization of his beliefs. But where we find such fallings-short, such imperfect and inappropriate proclivities and inactivities, we will *thereby* diminish our grounds for ascribing belief content to mechanisms we find.

It is unlikely, I have said, that the illata we eventually favor in academic psychology will resemble the putative illata of folk theory enough to tempt us to identify them. But whatever illata we find, we will interpret them and assign content to them by the light of our holistic attribution to the agent of beliefs and desires. We may not find structures in the agent that can be made to line up belief-by-belief with our intentional system catalog of beliefs for the agent. On Stich's view, and on Fodor's, we would be constrained to interpret this outcome – which all grant is possible – as the discovery that *there were no such things as beliefs after all*. Folk psychology was just false. On my view we would instead interpret this discovery – and a very likely one it is – as the discovery that the concrete systems of representations whereby brains realize intentional systems are simply not *sentential* in character.[11]

Of course sometimes there are sentences in our heads, which is hardly surprising, considering that we are language-using creatures. These sentences, though, are as much in need of interpretation via determination of our beliefs and

desires as are the public sentences we utter. Suppose the words occur to me (just "in my head"): 'Now is the time for violent revolution!' – did I thereby *think* the thought with the content that now is the time for violent revolution? It all depends, doesn't it? On what? On what I happened to believe and desire and intend when I internally uttered those words "to myself." Similarly, if "cerebroscopes" show that while the boy was handing me my change he was internally accompanying his transaction with the conscious or subconscious expression in his natural language or in Mentalese: 'this is the right change,' that would not settle the correct interpretation of that bit of internal language and hence would not settle the intentional interpretation of his act. And since he has made a mistake, there is no unqualified catalog of his intentional states and acts of the moment.

So I stick to my guns: even for the everyday cases of error Stich presents, the problems of belief-interpretation encountered by my view *really are there* in the folk-psychological practice, although they often lurk behind our confabulations and excuses. Nor will they go away for Stich's proposed alternative theory of content ascription. This is *not* to say that such phenomena cannot be given any coherent description. Of course they can be coherently described from either the design stance or the physical stance – a point on which Stich and I agree. So I do not discover any truths of folk theory I must regretfully forswear.