Making Sense of Ourselves

DANIEL C. DENNETT
Tufts University

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 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 idiosyncracies, shortcomings, and cognitive growth" (p. 48). 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 consitutes 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 acknowledgment 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.*) 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 the 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 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 foreswear.

* * *

In thus resisting Stich's objections, and keeping rationality at the foundation of belief and desire attribution, am I taking what Stich calls the *hard line*, or the *soft line*? The hard line, according to Stich, insists that intentional system theory's idealizing assumption of rationality is actually to be found in the folk practice from which intentional system theory is derived. The soft line "proposes some fiddling with the idealized notion of an intentional system" to bring it more in line with folk practice, which does not really (Stich insists) invoke considerations of rationality at all. These distinct lines are Stich's inventions, born of his frustration in the attempt to make sense of my expression of my view, which is both hard and soft—that is to say, flexible. The *flexible line* insists both that the assumption of rationality is to be found in the folk practice and that what rationality is is not what it appears to be to some theorists—so the idealization will require some "fiddling." What, then, do I say of the ideal of rationality exploited self-consciously by the intentional system strategist and as second nature by the rest of the folk?

Here Stich finds me faced with a dilemma. If I identify rationality with *logical consistency and deductive closure* (and the other dictates of the formal normative systems such as game theory and the calculus of probability) I am embarrassed by absurdities. Deductive closure, for instance, is just too strong a condition, as Stich's case of Oscar the engineer witnesses.[12] If, flying to the other extreme, I identify rationality with *whatever it is that evolution has provided us*, I either lapse into uninformative tautology or fly in the face of obvious counterexamples: cases of evolved manifest irrationality. What then do I say rationality is? I don't say.

Stich is right; for ten years I have hedged and hinted and entertained claims that I have later qualified or retracted. I didn't know what to say, and could see problems everywhere I turned. With that *mea culpa* behind me, I will now take the offensive, however, and give what I think are good reasons for cautiously resisting the demand for a declaration on the nature of rationality while still insisting that an assumption of rationality plays the crucial role I have seen for it.

First, a few words on what rationality is *not*. It is not deductive closure. In a passage Stich quotes from "Intentional Systems" I present the suggestion that "If S were ideally rational . . . S would believe every logical consequence of every belief (and ideally, S would have no false beliefs)" and I make a similar remark in "True Believers." That is, after all, the logically guaranteed resting point of the universally applicable, indefinitely extendable demand that one believe the "obvious" consequences of one's genuine, fully understood beliefs. But Stich's example of Oscar nicely reveals what is wrong with letting sheer entailment expand a rational agent's beliefs, and as Lawrence Powers shows in his important article "Knowledge by Deduction"[13] there is work to be done by a theory of knowledge *acquisition* by deduction: one *comes* to know (and believe) what one *didn't* already know (or believe) by deducing propositions from premisses already believed—a familiar and "obvious" idea, but one that requires the very careful exposition and defense Powers gives it. And it is important to note that in the course of making his case for what we might call implication–insulated cognitive states, Powers must advert to neologism and caveat: we must talk about what our agent "pseudo–believes" and "pseudo–knows" (p. 360ff). It puts one in mind, in fact, of Stich's own useful neologism for belief–like states lacking the logical fecundity of beliefs: "sub–doxastic states"[14].

Nor is rationality perfect logical consistency, although the *discovery* of a contradiction between propositions one is inclined to assent to is always, of course, an occasion for sounding the epistemic alarm.[15] Inconsistency, when discovered, is of course to be eliminated one way or another, but making the rooting out of inconsistency the pre–eminent goal of a cognizer would lead to swamping the cognitive system in bookkeeping and search operations to the exclusion of all other modes of activity.[16] Now how can I talk this way about inconsistency, given my account of the conditions for correct belief attribution? Who said anything about inconsistency of *beliefs*? When one enters the domain of considerations about the wise design of cognitive structures and operations, one has left belief proper behind, and is discussing, in effect, structurally identified features with more–or–less apt intentionalistic labels (see "Three Kinds of Intentional Psychology" and *Brainstorms*, pp. 26–27).

If I thus do not identify rationality with consistency and deductive closure, what then could be my standard? If I turn to evolutionary considerations, Stich suggests, "such established theories as deductive and inductive logic, decision theory and game theory" will be "of no help in assessing what an organism 'ought to believe'." This is just not true. The theorist who relinquishes the claim that these formalisms are the final *benchmark* of rationality can still turn to them for help, can still exploit them in the course of criticizing (on grounds of irrationality) and reformulating strategies, designs, interpretations. The analogy is imperfect, but just as one may seek help from a good dictionary, or a good grammar book, in supporting one's criticism of someone's spelling, word choice, or grammar, so may one appeal to the defeasible authority of, say, decision theory in objecting to someone's strategic formulation. One can also reject as wrong—or irrational—the advice one gets from a dictionary, a grammar, a logic, or any other normative theory, however well established.[17]

What of the evolutionary considerations? I am careful *not* to define rationality in terms of what evolution has given us—so I avoid outright tautology. Nevertheless, the relation I claim holds between rationality and evolution is more powerful than Stich will grant. I claim, as he notes, that if an organism is the product of natural selection we can assume that *most* of its beliefs will be true, and *most* of its belief–forming strategies will be rational. Stich disagrees: "it is simply not the case that natural selection favors true beliefs over false ones," because all natural selection favors is beliefs "that yield selective advantage" and "there are many environmental circumstances in which false beliefs will be more useful than true ones." I do not think it is *obvious* that it is *ever* advantageous to be designed to arrive at false beliefs about the world, but I have claimed that there are describable circumstances—rare circumstances—where it can happen, so I agree with Stich on this point: "*better safe than sorry* is a policy that recommends itself to natural selection," Stich says, echoing my claim in "Three Kinds of Intentional Psychology"—"Erring on the side of prudence is a well recognized good strategy, and so Nature can be expected to have valued it on occasions when it came up" (p. 45n).

But does this go any way at all toward rebutting my claim that natural selection guarantees that *most* of an organism's beliefs will be true, *most* of its strategies rational? I think not. Moreover, even if a strategy is, as I grant it very well may be, a "patently invalid" strategy that works most of the time in the contexts it is invoked—does this show it is an *irrational* strategy? Only if one is still clinging to the ideals of Intro Logic for one's model of rationality. It is not even that there are no "established" academic canons of rationality in opposition to the logicians' to which one might appeal. Herbert Simon is duly famous for maintaining that *it* *is rational* in many instances to *satisfice*—e.g., to leap to possibly "invalid" conclusions when the costs of further calculation probably outweigh the costs of getting the wrong answer. I think he is right, so I for one would not tie rationality to any canons that prohibited such practices. Stich declares:

> 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. [pp. 53–54]

This is a puzzling claim, for there are normative theories for different purposes, including the purposes of "generally getting the right answer." If one views these as at odds with one another, one makes a mistake. Deductive logic might be held to advise that in the face of uncertainty or lack of information one should simply *sit tight and infer nothing*—bad advice for a creature in a busy world, but fine advice if avoiding falsehood *at all costs* is the goal. It is better to recognize the various uses to which such strategies can be put, and let rationality consist in part of a good sense of when to rely on what. (It is also useful to remind ourselves that only a tiny fraction of all the "rational animals" that have ever lived have ever availed themselves self–consciously of *any* of the formal techniques of the normative theories that have been proposed.)

The concept of rationality is indeed a slippery concept. We agree, it seems, that a system would be improperly called irrational if although its *normal, designed* operation were impeccable (by the standards of the relevant norms), it suffered occasional *malfunctions*. But of course a system that was particularly delicate, particularly prone to uncorrected malfunctions, would hardly be a well–designed system; a system that was foolproof or failsafe would in this regard be better. But which would be better—which would be more rational—all things considered: a very slow but virtually failsafe system, or a very fast but only 90% malfunction–free system? It depends on the application, and there are even normative canons for evaluating such choices in some circumstances.

I want to use "rational" as a general–purpose term of cognitive approval—which requires maintaining only conditional and revisable allegiances between rationality, so considered, and the proposed (or even universally acclaimed) methods of getting ahead, cognitively, in the world. I take this usage of the term to be quite standard, and I take *appeals to* rationality by proponents of cognitive disciplines or practices to require this understanding of the notion. What, for instance, could Anderson and Belnap be appealing to, what could they be assuming about their audience, when they recommend their account of entailment over its rivals, if not to an assumably shared rationality which is such that it is an *open question* which formal system best captures it?[18] Or consider this commentary on the discovery that a compartmentalized memory is a necessary condition for effective cognition in a complex, time–pressured world:

> We can now appreciate both the costs and the benefits of this strategy; *prima facie*, the resulting behavior can be characterized as departures from rationality, but on the assumption that exhaustive memory search is not feasible, such memory organization is advisable overall, despite its costs. Correspondingly, a person's action may seem irrational when considered in isolation, but it may be rational when it is more broadly considered as part of the worthwhile price of good memory management.[19]

The claim is that it is rational to be inconsistent sometimes, not the pseudo–paradoxical claim that it is rational sometimes to be irrational. As the example shows, the concept of rationality is systematically pre–theoretical. One may, then, decline to *identify* rationality with the features of any formal system or the outcome of any process and still make appeals to the concept, and assertions about appeals to it (such as mine), without thereby shirking a duty of explicitness.

* * *

When one leans on our pre-theoretical concept of rationality, one relies on our shared intuitions—when they *are* shared, of course—about what makes sense. What else, in the end, could one rely on? What else would it be *rational* to rely on? When considering what we *ought to do*, our reflections lead us eventually to a consideration of what we *in fact do*; this is inescapable, for a catalogue of our considered intuitive judgments on what we ought to do is both a compendium of what we *do* think, and a shining example (by our lights—what else?) of how we *ought* to think.[20]

Now it will appear that I am backing into Stich's own view, the view that when we attribute beliefs and other intentional states to others, we do this by comparing them *to ourselves*, by projecting ourselves into their states of mind. One doesn't ask: "what ought this creature believe?" but "what would *I* believe if I were in its place?" (I have suggested to Stich that he call his view *idealogical solipsism*, but he apparently feels this would court confusion with some other doctrine.) Stich contrasts his view with mine and claims that "the notion of idealized rationality plays *no role at all*" [Stich's emphasis] in his account. "In ascribing content to belief states we measure others not against an idealized standard but against ourselves." But for the reasons just given, measuring "against ourselves" *is* measuring against an idealized standard.

Now Stich at one point observes that "since we take ourselves to approximate rationality, this explains the fact, noted by Dennett, that intentional description falters in the face of egregious irrationality." He must grant, then, that since we take ourselves to approximate rationality, it is also true that the results of his method and my method will coincide very closely. He, asking "what would I do if . . . ?" and I, asking "what ought he to do . . . ?" will typically arrive at the same account, since Stich will typically suppose that he would do what he ought to do, and I would typically suppose that what he ought to do is what I would do if I were in his shoes. If the methods were actually extensionally equivalent, one might well wonder about the point of the quarrel, but is there not room for the two methods to diverge in special cases? Let us see.

Can it be like this? Stich, cognizant of his lamentable and embarrassing tendency to affirm the consequent, imputes this same tendency to those whose beliefs and desires he is trying to fathom. He does this instead of supposing they might be free from his own particular foible, but guilty of others. Unlikely story. Here is a better one. Having learned about "cognitive dissonance," Stich is now prepared to find both in himself and in others the resolution of cognitive dissonance in the favoring of a self–justifying belief over a less comfortable belief better supported by the evidence. This is a fine example of the sort of empirical discovery that can be used to tune the intentional stance, by suggesting hypotheses to be tested by the attributer, but how would Stich say it had anything to do with *ourselves*, and how would this discovery be put into effective use independently of the idealizing assumption? For, first, is it not going to be an empirical question whether all people respond to cognitive dissonance as we do? If Stich builds this (apparently) sub–optimal proclivity into his very method of attribution, he foregoes the possibility of discovering varieties of believers happily immune to this pathology.

Moreover, consider how such an assumption of sub–optimality would get used in an actual case. Jones has just spent three months of hard work building an addition to his house; it looks terrible. Something must be done to resolve the uncomfortable cognitive dissonance. Count on Jones to slide into some belief that will save the situation. But which one? He might come to believe that the point of the project, really, was to learn all about carpentry by the relatively inexpensive expedient of building a cheap addition. Or he might come to believe that the bold thrust of the addition is just the touch that distinguishes his otherwise hackneyed if "tasteful" house from the run of the neighborhood houses. Or, . . . for many possible variations. But which of these is actually believed will be determined by seeing what he says and does, and then asking: what beliefs and desires would make those acts rational? And whatever delusion is embraced, it must be—and *will* be—carefully surrounded by plausible supporting material, generatable on the counterfactual assumption that the delusion is an entirely rationally held belief. Given what we already know about Jones, we might be able to predict which comforting delusion would be most attractive and efficient for him—that is, which would most easily cohere with the rest of the fabric of his beliefs. So even in a case of cognitive dissonance, where the beliefs we attribute are not optimal by anyone's lights, the test of rational coherence is the preponderant measure of our attributions.

I do not see how my method and Stich's can be shown to yield different results, but I also do not see that they could not. I am not clear enough about just what Stich is asserting. An interesting idea which is lurking in Stich's view is that when we interpret others we do so not so much by *theorizing* about them as by *using ourselves as analogue computers* that produce a result. Wanting to know more about *your* frame of mind, I somehow put myself in it, or as close to being in it as I can muster, and see what I thereupon think (want, do . . .).[21] There is much that is puzzling about such an idea. How can it work *without* being a kind of theorizing in the end? For the state I put myself in is not belief but make–believe belief. If I make believe I am a suspension bridge and wonder what I will do when the wind blows, what "comes to me" in my make–believe state depends on how sophisticated my knowledge is of the physics and engineering of suspension bridges. Why should my making believe I have your beliefs be any different? In both cases, knowledge of the imitated object is needed to drive the make–believe "simulation," and the knowledge must be organized into something rather like a theory.

Moreover, establishing that we do somehow arrive at our interpretations of others by something like simulation and self–observation would not by itself show that the guiding question of our effort is "what would I believe?" *as opposed to* "what ought he to believe?" A wary attributer might exhibit the difference by using the trick of empathy or make–believe to *generate* a candidate set of attributions to *test* against his "theory" of the other before settling on them. Note that the issue is far from clear even in the case of imagined *self*–attribution. What would your state of mind be if you were told you had three weeks to live? How do you think about this? In a variety of ways, probably; you do a bit of simulation and see what you'd say, think, and so on, and you also reflect on what kind of a person you think you are—so you can conclude that a person *like that* would believe—ought to believe—or want such–and–such.

Stich's paper raises many more problems well worth a response from me, but the deadline for this issue of *Philosophical Topics* mercifully intervenes at this point. I close with one final rejoinder. Stich seeks to embarrass me in closing with a series of rhetorical questions about what a frog *ought to believe*—for I have made my determination of what a frog *does* believe hinge on such questions. I grant that such questions are only problematically answerable under even the best conditions,[22] but view that as no embarrassment. I respond with a rhetorical question of my own: does Stich suppose that the exact content of what a frog does in fact believe is any more likely of determination?

NOTES

[1] See Stephen Stich's review of *Brainstorms*; "Headaches," *Philosophical Books*, April, 1980, and my reply, ibid.

[2] Wason and Johnson–Laird, and Nisbett and Ross (see Stich's notes 8 and 11). See also S. Stich and R. Nisbett, "Justification and the Psychology of Human Reasoning" in *Philosophy of Science*, 1980, Vol. 47, No. 2, pp. 188–202.

[3] See J. Weizenfeld "Surprise and Intentional Content," presented at the 3rd Annual meeting of the Society for Philosophy and Psychology, Pittsburgh, March 1977.

[4] Cf. Jerry Fodor, "Computation and Reduction" in C. W. Savage, ed., *Perception and Cognition: Issues in the Foundations of Psychology* , 1978, pp. 229–60.

[5] E.g., L. Jonathan Cohen, "Can Human Irrationality Be Experimentally Demonstrated?" forthcoming in *Behavioral and Brain Sciences*.

[6] I will continue to fly in the face of the examples raised by Vendler et al., about the differences between the objects of knowledge and the objects of belief until I can see that this imprecision is *dangerous*. Perhaps I will be shown this tomorrow, but I haven't been shown it yet.

[7] See C. Cherniak, "Rationality and the Structure of Human Memory" (Tufts University Cognitive Science Working Papers WP13, June 1980).

[8] Cf. Ryle, "A Puzzling Element in the Notion of Thinking" (1958), a British Academy Lecture reprinted in P. F. Strawson, ed., *Studies in the Philosophy of Thought and Action*, 1968, Oxford University Press.

[9] See, especially R. Nisbett and T. DeC. Wilson. "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review*, 1977.

[10] Michael Gazzaniga and J. E. Ledoux advocate a position along these lines in *The Integrated Mind* (New York: Plenum Press, 1978). For graphic accounts of confabulations in victims of brain disorders, see also Howard Gardner, *The Shattered Mind: the Person After Brain Damage*, Knopf, New York, 1975.

[11] See my "Beyond Belief" forthcoming in Andrew Woodfield, ed., *Thought and Object*, Oxford University Press, 1981.

[12] Cf. also Jerry Fodor, "Three Cheers for Propositional Attitudes" forthcoming in *Representations*, Bradford Books, 1981.

[13] *Philosophical Review*, July 1978, pp. 337–71.

[14] "Belief and Sub-Doxastic States," *Philosophy of Science*, December, 1978, pp. 499–518.

[15] See R. de Sousa, "How to Give a Piece of Your Mind; or The Logic of Belief and Assent," *Review of Metaphysics*, September 1971, pp. 52–79.

[16] See C. Cherniak, "Rationality and the Structure of Human Memory," op. cit., and Howard Darmstadter, "Consistency of Belief," *Journal of Philosophy*, May 20, 1971, pp. 301–10. The point has often been made in different contexts by Marvin Minsky as well.

[17] See, e.g., L. Jonathan Cohen, op. cit., and for a dissenting view, see S. Stich and R. Nisbett, "Justification and the Psychology of Human Reasoning," *Philosophy of Science*, June, 1980, pp. 188–202.

[18] A. R. Anderson and N. Belnap, *Entailment: the Logic of Relevance and Necessity*, (Princeton: Princeton University Press, 1974).

[19] Cherniak, op. cit., p. 23.

[20] "Thus, what and how we do think is evidence for the principles of rationality, what and how we ought to think. This itself is a methodological principle of rationality; call it the *Factunorm Principle*. We are (implicitly) accepting the Factunorm Principle whenever we try to determine what or how we ought to think. For we must, in that very attempt, think. And unless we can think that what and how we do think there is correct—and thus is evidence for what and how we ought to think—we cannot determine what or how we ought to think." R. Wertheimer, "Philosophy on Humanity," in R. L. Perkins, ed., *Abortion: Pro and Con*, Schenkman, Cambridge, Mass., 1974, p. 110–111. See also Nelson Goodman, *Fact, Fiction, and Forecast*, 2nd edition, 1965, p. 63.

[21] Adam Morton's new book *Frames of Mind* (Oxford University Press, 1980) has much to say on this topic which I have not yet had an opportunity to digest. Hence my tentative and sketchy remarks on this occasion.

[22] Cf. Dennett, *Content and Consciousness* (London: 1969, Routledge & Kegan Paul), pp. 83–85.