STEPHEN P. STICH

COULD MAN BE AN IRRATIONAL ANIMAL?

Some Notes on the Epistemology of Rationality

1.

Aristotle thought man was a rational animal. From his time to ours, however, there has been a steady stream of writers who have dissented from this sanguine assessment. For Bacon, Hume, Freud, or D. H. Lawrence, rationality is at best a sometimes thing. On their view, episodes of rational inference and action are scattered beacons on the irrational coastline of human history. During the last decade or so, these impressionistic chroniclers of man's cognitive foibles have been joined by a growing group of experimental psychologists who are subjecting human reasoning to careful empirical scrutiny. Much of what they have found would appall Aristotle. Human subjects, it would appear, regularly and systematically invoke inferential and judgmental strategies ranging from the merely invalid to the genuinely bizarre.

Recently, however, there have been rumblings of a reaction brewing – a resurgence of Aristotelian optimism. Those defending the sullied name of human reason have been philosophers, and their weapons have been conceptual analysis and epistemological argument. The central thrust of their defense is the claim that empirical evidence could not possibly support the conclusion that people are systematically irrational. And thus the experiments which allegedly show that they are must be either flawed or misinterpreted.

In this paper I propose to take a critical look at these philosophical defenses of rationality. My sympathies, I should note straightaway, are squarely with the psychologists. My central thesis is that the philosophical arguments aimed at showing irrationality cannot be experimentally demonstrated are mistaken. Before considering these arguments, however, we would do well to set out a few illustrations of the sort of empirical studies which allegedly show that people depart from normative standards of rationality in systematic ways. This is the chore that will occupy us in the following section.

Synthese **64** (1985) 115–135. 0039–7857/85.10

2.

One of the most extensively investigated examples of inferential failure is the so-called "selection task" studied by P. C. Wason, P. N. Johnson-Laird, and their colleagues (1970, 1977, 1972, Chaps. 13–15). A typical selection task experiment presents subjects with four cards like those in Figure 1. Half of each card is masked. Subjects are then given the following instructions:

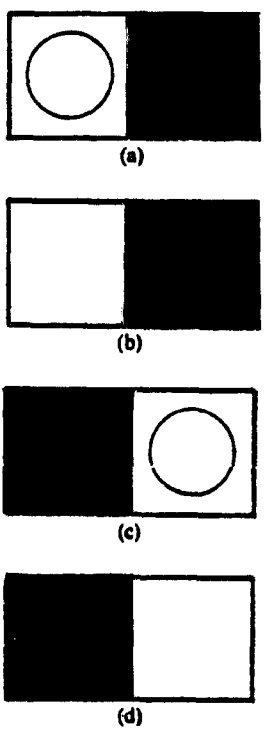

Fig. 1.

Which of the hidden parts of these cards do you need to see in order to answer the following question decisively?
FOR THESE CARDS IS IT TRUE THAT IF THERE IS A CIRCLE ON THE LEFT THERE IS A CIRCLE ON THE RIGHT?
You have only one opportunity to make this decision; you must not

assume that you can inspect cards one at a time. Name those cards which it is absolutely essential to see.

Wason and Johnson-Laird discovered that subjects, including very intelligent subjects, find the problem remarkably difficult. In one group of 128 university students, only *five* got the right answer. Moreover, the mistakes that subjects make are not randomly distributed. The two most common wrong answers are that one must see both (a) and (c), and that one need only see (a). The phenomenon turns out to be a remarkably robust one, producing essentially the same results despite significant variation in the experimental design, the wording of the question and the details of the problem. For example, subjects presented with the four envelopes in Figure 2 and asked which must be turned over to

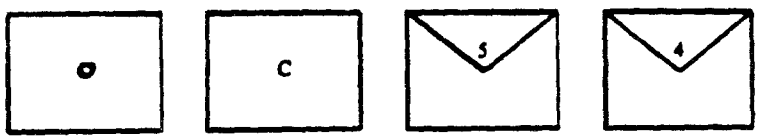

Fig. 2.

determine the truth of the rule:

IF IT HAS A VOWEL ON ONE SIDE IT HAS AN EVEN NUMBER ON THE OTHER

do just as badly as subjects given the cards in Figure 1. However, there are variations in the experimental design which substantially improve inferential performance. One of these is making the relation between the antecedent and the consequent of the conditional rule in the instructions more "realistic". So, for example, subjects presented with the envelopes in Figure 3, and asked which must be turned over to

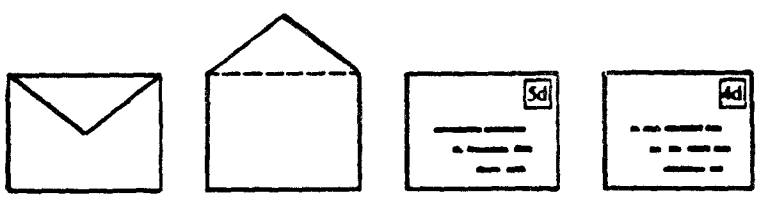

Fig. 3.

determine the truth of the rule:

IF IT IS SEALED, THEN IT HAS A 5d STAMP ON IT

do vastly better than subjects presented with the envelope in Figure 2. In one experiment using the "realistic" material, 22 out of 24 subjects got the right answer.[1]

Wason and Johnson-Laird have also explored the ways in which subjects react when they are shown that their initial inferences are mistaken. In Figure 1, for example, a subject who said he must see only the hidden side of (a) might be asked to remove the masks on both (a) and (d), discovering a circle under each mask. Many subjects have a startling reaction. They note that the rule is false for these cards – in virtue of card (d) – and they continue to insist that it was only necessary to see card (a)! In further work Wason, Johnson-Laird and their colleagues have looked at the ways in which subjects react when the apparent contradiction in their claims is pointed out. The intriguing details of these studies need not detain us here.

My second example of research revealing *prima facie* deviation from normative standards of inference focuses on the way people assess the probability of logically compound events or states of affairs. It is a truism of probability theory that the likelihood of a compound event or state of affairs must be less than or equal to the likelihood of the component events or states of affairs. If the components are probabilistically independent, the probability of the compound is equal to the product of the probabilities of the components. If the components are not probabilistically independent, matters are more complicated. But in no case will the probability of the compound be *greater* than the probability of the components. There are, however, a number of experiments which demonstrate that people regularly violate this basic tenet of probabilistic reasoning. In one such experiment Kahneman and Tversky gave subjects personality profiles of various target persons. Subjects were then asked to assess the likelihood that the persons described in the profiles belonged to various groups. One group of subjects was asked to estimate the likelihood that profiled persons were members of noncompound groups like *lawyers* or *republicans*. Another group of subjects was asked to estimate the probability that the profiled persons were members of compound groups like *republican lawyers*. What Tversky and Kahneman (1982) found is that if a

profiled person is judged rather unlikely to be, say, a lawyer, and rather likely to be a Republican, he will be judged moderately likely to be a Republican lawyer. This is, the likelihood of the target being a Republican lawyer is judged significantly higher than the likelihood of his being a lawyer! The explanation that Kahneman and Tversky offer for these peculiar judgments turns on what they call the representativeness heuristic. Subjects, they hypothesize, assess the likelihood that a target person is a Republican lawyer by assessing the similarity between the profile and the stereotypical Republican, assessing the similarity between the profile and the stereotypical lawyer, and then *averaging* these two likelihoods.

In a similar study with alarming implications for public policy judgments, Slovic, Fischoff, and Lichtenstein (1977) showed that subjects estimate the probability of a compound sequence of events to be greater than the least likely of the events in the sequence. It is disquieting to speculate on how large an impact this inferential failing may have on people's assessments of the chance of such catastrophes as nuclear reactor failures which require a number of distinct events to occur in sequence (Slovic and Fischoff, 1978).

My final example of an experimental program exploring human irrationality is the work on belief perseverance by Ross, Lepper, and their colleagues (1975). One of the experimental strategies used in this work is the so-called "debriefing" paradigm. In these experiments subjects are given evidence which is later completely discredited. But despite being "debriefed" and told exactly how they had been duped, subjects tend to retain to a substantial degree the beliefs they formed on the basis of the discredited evidence. In one such experiment subjects were presented with a task of distinguishing between authentic and unauthentic suicide notes. As they worked they were provided with false feedback indicating that overall they were performing at close to the average level or (for other subjects) much above the average level, or (for a third group of subjects) much below the average level. Following this, each subject was debriefed, and the predetermined nature of the feedback was explained to him. They were not only told that their feedback had been false but were also shown the experimenter's instruction sheet assigning them to the success, failure, or average group, and specifying the feedback to be presented. Subsequent to this, and allegedly for quite a different reason, subjects were asked to fill out a questionnaire on which they were asked to estimate their actual

performance at the suicide note task, to predict their probable success on related future tasks and to rate their ability at suicide note discrimination and other related tasks. The results revealed that even after debriefing subjects who had initially been assigned to the success group continued to rate their performance and abilities far more favorably than did subjects in the average group. Subjects initially assigned to the failure group showed the opposite pattern of results. Once again, these results appear to reflect a robust phenomenon which manifests itself in many variations on the experimental theme, including some conducted outside the laboratory setting.

The three examples I have sketched could easily be supplemented by dozens more, all apparently demonstrating that human reasoning often deviates substantially from the standard provided by normative canons of inference. Let us now turn our attention to the arguments aimed at showing that these experiments are being misinterpreted.

3.

Of the three arguments I shall consider, two are due to D. C. Dennett. Both arguments are embedded in Dennett's much more elaborate theory about the nature of intentional attributions, though neither argument is developed in much detail. In a pair of previous papers (Stich, 1980, 1981a) I have tried to give a systematic critique of Dennett's views with due attention to problems of interpretation and the possibilities of alternative construals. In the present paper I will sidestep most of these niceties. What I wish to show is that a pair of arguments are mistaken. I think it is clear that Dennett has at least flirted with each of these arguments. But for the purposes at hand, pinning the tail on the donkey is of little importance.

The first of the arguments I am attributing to Dennett might be called *the argument from the inevitable rationality of believers*. On Dennett's view, when we attribute beliefs, desires, and other states of common sense psychology to a person, or for that matter to an animal or an artifact, we are assuming or presupposing that the person or object can be treated as what Dennett calls an *intentional system*. An intentional system is one which is rational through and through; its beliefs are "those it ought to have, given its perceptual capacities, its epistemic needs, and its biography.... Its desires are those it ought to have, given its biological needs and the most practicable means of satisfying them....

And its behavior will consist of those acts that it *would be rational* for an agent with those beliefs and desires to perform." (1981a) According to Dennett it is in the context of this set of assumptions about rationality that our ordinary talk about beliefs, desires, or other intentional states gains its meaning. If this is right, then we should expect that when a person's behavior is less than fully rational the intentional scheme would no longer apply. We could not rest content with a description of a person as holding an incoherent or irrational set of beliefs, for if rationality is absent, we cannot coherently ascribe beliefs at all. Dennett (1978, p. 20) puts the matter as follows:

> 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.

Given this much of Dennett's view, it follows straightforwardly that no experiment could demonstrate that people systematically invoke invalid or irrational inferential strategies. The point is not that people *must* be rational. No such conclusion follows from Dennett's view. What does follow from Dennett's view is that people must be rational *if they can usefully be viewed as having any beliefs at all.* We have no guarantee that people will behave in a way that makes it profitable for us to assume the intentional stance toward them. But intentional descriptions and rationality come in the same package; there is no getting one without the other. Thus if people infer at all, that is, if they generate new beliefs from old ones, from perceptual experience, or what have you, then they must do so rationally. Dennett is, in effect, offering us a *reductio* on the claim that people infer irrationally. If a system infers irrationally, it cannot be an intentional system; thus we cannot ascribe beliefs and desires to it. But since inference is a belief generating process, the system does not infer at all.

Now as I see it, the problem with Dennett's argument comes right at the beginning. He is simply wrong about the relationship between our ordinary notions of belief and desire and his notion of an idealized fully rational intentional system. *Pace* Dennett, it is simply not the case that

our ordinary belief and desire ascriptions presuppose full rationality. There is nothing in the least incoherent or unstable about a description, cast in intentional terms, of a person who has inconsistent beliefs. The subjects in Wason and Johnson-Laird's experiments provide a clear example, one among endlessly many. Some of these subjects clearly believe that cards (a) and (c) must be removed, and defend their view with considerable vigor. Yet these subjects clearly understand the conditions of the problem and have no false beliefs about what they are being asked to do.[2]

In defending his contention that ordinary intentional ascriptions gain their meaning against the background of a theory of intentional systems, Dennett offers a pair of arguments, one long and one short. The short one is the observation, attributed to Quine, that blatant or obvious inconsistency is the best evidence we can have that we are misdescribing a subject's beliefs. This fact is readily explained if belief ascription presupposes full rationality. The longer argument has much the same structure. In effect, Dennett maintains that his intentional system explication of ordinary belief and desire talk explains many of the facts about the way we use these locutions in describing and explaining the behavior of persons, animals, and artifacts. All of this I cheerfully grant. I also grant that, until recently at least, Dennett's explication of ordinary intentional locutions was the best – indeed pretty near the only – game in town. None of this, however, persuades me to accept Dennett's explication. The reason is that I think there is a better explication of the way we use our workaday belief and desire locutions, an explication that handles all the facts Dennett's can handle without the paradoxical consequence that intentional descriptions of irrational beliefs are unstable or incoherent. The basic idea of this alternative explication is that, in using intentional locutions we are presupposing that the person or system to which they are applied is, in relevant ways, similar to ourselves. Thus inferential errors that we can imagine ourselves making – errors like those recounted in my previous section – can be described comfortably in intentional terms. It is only the sort of error or incoherence that we cannot imagine falling into ourselves that undermines intentional description. This is the reason that blatant inconsistency of the sort Quine has in mind is evidence that something has gone wrong in our intentional attributions. Plainly the alternative "similar-to-us" account of intentional locutions needs a much more detailed elaboration. I have made a beginning at this in Stich (1981b).[3]

4.

Dennett concedes that his second argument is uncomfortably vague, so a fair bit of interpretation is needed. I will call this one *the argument from natural selection*. The closest Dennett comes to settling out the argument is in a passage where he reflects on whether we could adopt the intentional stance toward thoroughly exotic creatures encountered on another planet. His answer is that we could, provided "we have reason to suppose that a process of natural selection has been in effect." But why would the mere existence of natural selection suffice to insure that the creatures would be good approximations to the thoroughly rational ideal embodied in the notion of an intentional system? Dennett offers no detailed answer, but provides us with a few hints, as have other writers who have sounded similar themes. These hints may be elaborated into the following argument.

1. Natural selection will favor (i.e., select for) inferential strategies which generally yield true beliefs. This is because, in general, true beliefs are more adaptive than false ones; they enable the organism to cope better with its environment. There are exceptions, of course. But on the whole organisms will outcompete their conspecifics if their ratio of true beliefs to false ones is higher. After an extended period of natural selection we can expect that the inferential strategies an organism uses will be ones which generally yield true beliefs.

2. An inferential strategy which generally yields true beliefs is a rational inferential strategy. Therefore,

3. Natural selection will favor rational inferential strategies.

Since Dennett's Martians are, ex hypothesis, the product of an extended process of natural selection we can conclude that they use rational inferential strategies. And, closer to home, since human beings are the result of millions of years of natural selection we know that they too must use rational inferential strategies. Thus any research program which claims to have evidence for widespread and systematic irrationality among humans must be misinterpreting its results. It is my suspicion that many writers who have recently been urging a naturalized or evolutionary reinterpretation of epistemology have had something very like this argument in mind. If so, then it is all the more important to focus critical scrutiny on the argument, for such scrutiny shows the argument to be seriously flawed.

Consider the first step. Is it true that natural selection favors inferential strategies which generally yield true beliefs? The answer, I

think, is clearly no. Perhaps the most vivid way to make the point is with a brief description of some intriguing experiments by John Garcia and his co-workers (1972). In one series of experiments Garcia's group fed rats distinctively flavored water or food, and then subjected them to substantial doses of radiation, enough to induce radiation sickness. After a single episode, the rats developed a strong aversion to the distinctively flavored food or water that had been used. Workers in other laboratories have demonstrated that the same phenomenon occurs even when the rat is exposed to radiation as much as 12 hours after eating or drinking. It has also been shown that the taste of the food is the object of the rat's aversion. The rats acquire no aversion to the cage in which the distinctive food was eaten, nor do they acquire an aversion to food pellets of a distinctive size. But if two substances are eaten in sequence prior to illness, novelty is a much more potent factor than recency in determination of the aversion. In short, the rat behaves as though it believes that anything which tastes like the distinctive tasting stuff it has eaten will cause it to become deathly ill. Moreover, it is clear that this belief, if that is what it is, is the result of an innate belief (or aversion) forming strategy which is surely the result of natural selection.

Consider now how often the inferential strategy which leads to the rat's belief will lead to a true belief? In the laboratory, of course, the inferential strategy is thoroughly unreliable. It is the radiation, not the food, which causes the rat's illness. But what about the rats in their natural environment? I know of no studies of rat epidemiology which indicate the most common causes of acute illness among rats. I would suspect, however, that rats, like people, fall victim to all manner of acute afflictions caused by viruses and bacteria which are not transmitted through food, still less through distinctively flavored food. If this is right, if, to be more specific, more than half of the illnessses rats endure in the wild which lead to the development of Garcia aversions are not transmitted by distinctively flavored food, it follows that *most* of the beliefs produced by the innate inferential strategy Garcia discovered are *false* beliefs. So it is just not true that natural selection favors inferential strategies which generally yield true beliefs. It is important to note that this argument does not turn essentially on my conjecture about the percentage of rat illnesses caused by distinctive tasting food. The real point of my argument is that *if* my conjecture is correct, it would pose no puzzle for the student of natural selection.

Natural selection might perfectly well opt for an inferential strategy which produces false beliefs more often than true ones. The sole concern of natural selection is with reproductive success and those features that foster it. When it comes to food poisoning, natural selection may well prefer an extremely cautious inferential strategy which is very often wrong, to a less cautious one which more often gets the right answer. It might be protested that the Garcia phenomenon does not really join the issue of irrational inference since the rats acquire an aversion, and aversions are not plausibly treated as beliefs. But this reply misses the essential point. Natural selection *could* perfectly well lead to inferential strategies which generally get the wrong answer, but are right when it counts most, just as it leads to aversions to foods most of which are harmless and nourishing. Often it is more adaptive to be safe than sorry.

Thus far my critique of the argument from natural selection has been aimed at the first step, the one which claims that natural selection favors inferential strategies that generally yield true beliefs. But even if we were to grant this dubious claim, the argument from natural selection would still be defective. For its second premise is false as well. That premise, recall, is that inferential strategies which generally yield the right answer are rational inferential strategies. In many cases this simply is not so. Perhaps the clearest examples of generally truth generating inferential strategies which are not rational are the cases in which a strategy is being invoked in a domain or setting significantly different from the one in which it presumably evolved. Once again an example from the study of animal behavior provides a striking illustration. Alcock (1975) recounts that a certain species of toad is capable of learning on a single trial to avoid eating a noxious species of millipede. However, the very same toad will continue to consume BBs that are rolled past it until it quite literally becomes a living beanbag! With only a bit of anthropomorphism, we might describe the case as follows. On seeing a millipede of a species previously found to be noxious, the toad comes to believe (i.e., infers) that it is not good to eat. But BBs, with their bland flavor, produce no such belief. Each time a new BB is rolled by, the toad infers that it is good to eat. This belief, of course, is quite false, a fact which will become obvious the first time the BB-filled toad attempts to leap out of harm's way. But, of course, the inferential strategy which lead to the belief *generally* yields true beliefs. Does this show that the strategy is normatively appropriate for the toad to use on

the BBs? I am inclined to think that the answer is no.

For all its vividness, the toad example may not be the best one to make my point. For some would protest that they just don't know what counts as a rational inferential strategy for a toad, a protest with which I have considerable sympathy. But the moral I want to draw from the toad example is one which can be drawn also from many cases involving human inference. A common theme in the research on human inference is that people are inclined to overextend the domain of an inferential strategy, applying it to cases where it is normatively inappropriate. Nisbett and Wilson (1977), for example, suggest that many causal inferences are influenced by a primitive version of the representativeness heuristic.

> People have strong *a priori* notions of the types of causes that ought to be linked to particular types of effects, and the simple "resemblance criterion" often figures heavily in such notions. Thus, people believe that great events ought to have great causes, complex events ought to have complex causes, and emotionally relevant events ought to have emotionally relevant causes. . . . The resemblance criterion is transparently operative in the magical thinking of prescientific cultures. For example Evans-Prichard . . . reported such Azande beliefs as the theory that fowl excrement was a cure for ringworm and the theory that burnt skull of red bush-monkey was an effective treatment for epilepsy. Westerners unacquainted with Azande ecology might be tempted to guess that such treatments were the product of trial and error or laboriously accumulated folk wisdom. Unfortunately, the truth is probably less flattering to Azande medical science. Fowl excrement resembles ringworm infection; the jerky, frenetic movements of the bush-monkey resemble the convulsive movements that occur during an epileptic seizure. (Nisbett and Ross, 1980, pp. 115–116).

Now it may well be that in a sufficiently primitive setting the primitive representativeness heuristic generally does get the right answer; it may have served our hunter-gatherer forebears in good stead. But it seems clear that the Azande are invoking the strategy in a domain where its applicability is, to say the least, normatively dubious. Nisbett and Ross go on to argue that the primitive representativeness heuristic plays a central role in psychoanalytic inference and in contemporary lay inference about the causes of disease, crime, success, etc. The normative inappropriateness of the heuristic in these settings is, I should think, beyond dispute.

The primitive representativeness heuristic is an extreme example of the overextension of an inferential strategy. For we have to go a long way back into our hunter-gatherer ancestry before coming upon life situations in which the heuristic is generally reliable and adaptive. But

many of the other inferential failings recounted in the recent literature would seem to arise in a similar way. An inference pattern which generally gets the right answer in a limited domain is applied outside that domain, often to problems without precedent during the vast stretches of human and pre-human history when our cognitive apparatus evolved. Indeed, it is disquieting to reflect on how vast a gap there likely is between the inferences that are important to modern science and society and those that were important to our prehistoric forebears. As Einstein noted, "the most incomprehensible thing about the universe is that it is comprehensible."[4]

I have been arguing that inferential strategies which generally get the right answer may nonetheless be irrational or normatively inappropriate when applied outside the problem domain for which they were shaped by natural selection. If this is right, then the second premise of the argument from natural selection must be rejected. Before leaving this topic I want to digress briefly to raise a thornier issue about normatively appropriate inference. It seems beyond dispute that an inferential strategy like the primitive representativeness heuristic is out of place in modern inquiries about the causes of cancer or of reactor failures. But what about the use of these heuristics in their natural settings? Are they normatively appropriate in those domains to which natural selection has molded them and in which (let us assume) they generally do produce the right answer? If I understand Professor Goldman's view correctly, he would answer with an unqualified affirmative. But I am less confident. At issue here is the deep and difficult question of just what we are saying of an inferential strategy when we judge that it is or is not normatively appropriate. This issue will loom large in the remaining pages of this paper.

Before leaving the argument from natural selection, we would do well to note one account of what it is for an inference strategy to be rational or normatively appropriate which had best be avoided. This is the reading which turns the conclusion of the argument from natural selection into a tautology by the simple expedient of defining *rational inferential strategy* as *inferential strategy favored by natural selection.* Quite apart from its *prima facie* implausibility, this curious account of rationality surely misses the point of psychological studies of reasoning. These studies are aimed at showing that people regularly violate the normative canons of deductive and inductive logic, probability theory, decision theory, etc. They do not aim at showing that people use

inferential strategies which have not evolved by natural selection!

5.

The final argument I want to consider is one proposed by L. Jonathan Cohen (1981). Cohen's argument grows out of an account of how we establish or validate normative theses about cognitive procedures – how we justify claims about rational or irrational inference. On Cohen's view normative theses about cognitive procedures are justified by what in ethics has come to be known as the method of *reflective equilibrium*. The basic input to the method, the data if you will, are intuitions, which Cohen characterizes as "immediate and untutored inclinations... to judge that" something is the case. In ethics the relevant intuitions are judgments about how people ought or ought not to behave. In the normative theory of reasoning they are judgments about how people ought or ought not to reason.

According to Cohen, a normative theory of reasoning is simply an idealized theory built on the data of people's individualized intuitions about reasoning. As in science, we build our theory so as to capture the bulk of the data in the simplest way possible. Our theory, in the case at hand, will be an interlocking set of normative principles of reasoning which should entail most individualized intuitions about how we should reason in the domain in question. An idealized theory need not aim at capturing all the relevant intuitions of all normal adults. Scattered exceptions – intuitions that are not entailed by the theory – can be tolerated in the same spirit that we tolerate exceptions to the predictions of the ideal gas laws.

Cohen stresses that normative theories of reasoning are not theories about the data (that is, about intuitions) any more than physics is a theory about observed meter readings, or ethics a theory about intuitions of rightness and wrongness. Just what normative theories *are* about is a question Cohen sidesteps.

> Fortunately, it is not necessary for present purposes to determine what exactly the study of moral value, probability or deducibility has as its proper subject matter. For example, an applied logician's proper aim may be to limn the formal consequences of linguistic definitions..., the most general features of reality... or the structure of ideally rational beliefs systems.... But, whatever the ontological concerns of applied logicians, they have to draw their evidential data from intuitions in concrete, individual cases; and the same is true for investigations into the norms of everyday probabilistic reasoning. (321)

But although a normative theory of reasoning is not a theory about reasoning intuitions, it is perfectly possible, on Cohen's view, to construct an empirical theory which is concerned to describe or predict the intuitive judgments which provide the data for the corresponding normative theory. This second theory

> will be a psychological theory, not a logical . . . one. It will describe a competence that human beings have – an ability, uniformly operative under ideal conditions and often under others, to form intuitive judgements about particular instances of . . . right or wrong, deducibility or nondeducibility, probability or improbability. This theory will be just as idealized as the normative theory . . . . (321)

Having said this much, Cohen can now neatly complete his argument for the inevitable rationality of normal people. The essential point is that the empirical theory of human reasoning, that is, the psychological theory that aims to describe and predict intuitive judgments, exploits the same data as the normative theory of reasoning, and exploits them in the same way. In both cases, the goal is to construct the simplest and most powerful set of principles that accounts for the bulk of the data. Thus, once a normative theory is at hand, the empirical theory of reasoning competence will be free for the asking, since it will be *identical* with the normative theory of reasoning! Though the empirical theory of reasoning competence "is a contribution to the psychology of cognition", Cohen writes,

> it is a by-product of the logical or philosophical analysis of norms rather than something that experimentally oriented psychologists need to devote effort to constructing. It is not only all the theory of competence that is needed in its area. It is also all that is possible, since a different competence, if it actually existed, would just generate evidence that called for a revision of the corresponding normative theory.
>
> In other words, where you accept that a normative theory has to be based ultimately on the data of human intuition, you are committed to the acceptance of human rationality as a matter of fact in that area, in the sense that it must be correct to ascribe to normal human beings a cognitive competence – however often faulted in performance – that corresponds by point with the normative theory. (321)

It is important to see that Cohen's view does not entail that people never reason badly. He can and does happily acknowledge that people make inferential errors of many sorts and under many circumstances. But he insists that these errors are performance errors, reflecting nothing about the underlying, normatively unimpeachable competence. The account Cohen would give of inferential errors is analogous to the account a Chomskian would give about the errors a person might make

in speaking or understanding his own language. We often utter sentences which are ungrammatical in our own dialect, but this is no reflection on our underlying linguistic competence. On the Chomskian view, our competence consists in a tacitly internalized set of rules which determines the strings of words that are grammatical in our language, and these rules generate no grammatical strings. Our utilization of these rules is subject to a whole host of potential misadventures which may lead us to utter ungrammatical sentences: there are slips of the tongue, failures of memory, lapses of attention, and no doubt many more. It is certainly possible to study these failures and thereby to learn something about the way the mind exploits its underlying competence. But while such studies might reveal interesting defects in performance, they could not reveal defects in competence. Analogously, we may expect all sorts of defects in inferential performance, due to inattention, memory limitations, or what have you. Study of these failings may indicate something interesting about the way we exploit our underlying cognitive competence. But such a study could no more reveal an irrational or defective cognitive competence than a study of grammatical errors could reveal that the speaker's linguistic competence was defective.

This is all I shall have to say by way of setting out Cohen's clever argument. As I see it, the argument comes to grief in the account it offers of the justification of normative theses about cognitive procedures. Perhaps the clearest way to underscore the problem with Cohen's epistemological account is to pursue the analogy between grammar and the empirical or descriptive theory of reasoning competence. Both theories are based on the data of intuition and both are idealized. But on Cohen's account there is one striking and paradoxical dis-analogy. In grammar we expect different people to have different underlying competences which manifest themselves in significantly different linguistics intuitions. The linguistic competence of a Frenchman differs radically from the linguistic competence of an Englishman, and both differ radically from the linguistic competence of a Korean. Less radical, but still significant, are the differences between the competence of an Alabama sharecropper, an Oxford don, and a Shetland Island crofter. Yet on Cohen's account of the empirical theory of reasoning there is no mention of different people having different idealized competences. Rather, he seems to assume that in the domain of reasoning all people have exactly the same competence. But why should we not expect that cognitive competence will vary just as much

as linguistic competence? The only answer I can find in Cohen's writing is a brief suggestion that cognitive competence may be *innate*. Yet surely this suggestion is entirely gratuitous. Whether or not individuals, social groups, or cultures differ in their cognitive competence is an *empirical* question, on all fours with the parallel question about linguistic competence. It is a question to be settled by the facts about intuitions and practice, not by a priori philosophical argument. And while the facts are certainly far from all being in, I am inclined to think that studies like those reviewed at the beginning of this paper, along with hundreds of others that might have been mentioned, make it extremely plausible that there are substantial individual differences in cognitive competence.

Now if this is right, if different people have quite different cognitive competences, then Cohen's account of the justification of a *normative* theory of reasoning faces some embarrassment. For recall that on this account a normative theory of reasoning is identical with a descriptive theory of cognitive competence; they are built on the same data and idealized in the same way. So if there are *many* cognitive competences abroad in our society and others, then there are *many* normative theories of cognition. But if there are many normative theories of cognition, which is the right one? Note that just here the analogy between linguistic competence and cognitive competence breaks down in an illuminating way. For although there are obviously great variations in linguistic competence, there is no such thing as normative theory of linguistics (or at least none that deserves to be taken seriously). Thus there is no problem about which of the many linguistic competences abroad in the world corresponds to the normatively correct one.

The problem I have been posing for Cohen is analogous to a familiar problem in ethics. For there too there is good reason to suspect that the method of reflective equilibrium would yield different normative theories for different people, and we are left with the problem of saying which normative theory is the right one. One response to the problem in ethics, though to my mind an unsatisfactory one, is a thoroughgoing relativism: my normative theory is the right one *for me*, yours is the right one *for you*. One way for Cohen to deal with the problem of the multiplicity of normative theories of cognition might be to adopt an analogous relativism. My inferential competence is right for me, yours is right for you. But this move is even more unpalatable for the

normative theory of cognition than it is for ethics. We are not in the least inclined to say that any old inference is normatively acceptable for a subject merely because it accords with the rules which constitute his cognitive competence. If the inference is stupid or irrational, and if it accords with the subject's cognitive competence, then his competence is stupid or irrational too, in this quarter at least.

A second strategy for dealing with the multiplicity of normative theories might be to adopt a majoritarian view according to which it is the cognitive competence of the majority that is normatively correct. This is no more plausible than the relativist alternative, however. First, it is not at all clear that there is a majority cognitive competence, any more than there is a majority linguistic competence. It may well be that many significantly different competences co-exist in the world, with the most common having no more than a meagre plurality. Moreover, even if there is a majority cognitive competence, there is little inclination to insist that it must be the normatively correct one. If, as seems very likely, most people disregard the impact of regression in estimating the likelihood of events, then most infer badly! (Nisbett and Ross, 1980, pp. 150 ff.).

The upshot of these reflections is that Cohen has simply told the wrong story about the justification of normative theories of cognition. Given the possibility of alternative cognitive competences, he has failed to tell us which one is normatively correct. Should he supplement his story along either relativist or majoritarian lines he would be stuck with the unhappy conclusion that a patently irrational inferential strategy might turn out to be the normatively correct one.[5]

By way of conclusion, let me note that there is a variation on Cohen's reflective equilibrium story which does a much better job of making sense of our normative judgments about reasoning, both in everyday life and in the psychology laboratory. It seems clear that we do criticize the reasoning of others, and we are not in the least swayed by the fact that the principles underlying a subject's faulty reasoning are a part of his – or most people's – cognitive competence. We are, however, swayed to find that the inference at hand is sanctioned or rejected by the cognitive competences of experts in the field of reasoning in question. Many well-educated people find statistical inferences involving regression to the mean to be highly counter-intuitive, at least initially. But sensible people come to distrust their own intuition on the matter when they learn that principles requiring regressive inference

are sanctioned by the reflective equilibrium of experts in statistical reasoning. In an earlier paper, Nisbett and I (1980) tried to parlay this observation into a general account of what it is for a normative principle of reasoning to be justified. On our view, when we judge someone's inference to be normatively inappropriate, we are comparing it to (what we take to be) the applicable principles of inference sanctioned by expert reflective equilibrium. On this account, there is no puzzle or paradox implicit in the practice of psychologists who probe human irrationality. They are evaluating the inferential practice of their subjects by the sophisticated and evolving standard of expert competence. From this perspective, it is not all that surprising that lay practice has been found to be markedly defective in many areas. We would expect the same, and for the same reason, if we examined lay competence in physics or in economics.

There is a hopeful moral embedded in this last observation. If, as Cohen suggests, cognitive competence is innate, then normatively inappropriate competence is ominous and inalterable. But if, as I have been urging, there is every reason to think that cognitive competence, like linguistic competence, is to a significant extent acquired and variable, then there is reason to hope that competence can be improved through education and practice, much as a child from Liverpool can acquire the crisp linguistic competence of an Oxford don. There is an important disanalogy, of course. Liverpudlean cadances are harmless and charming; normatively defective inference is neither. I am inclined to think it a singular virtue of recent studies of reasoning that they point to the areas where remedial education is needed most.

NOTES

[1] Johnson-Laird, Legrenzi and Sonino-Legrenzi (1972). However, see also Griggs and Cox (forthcoming).

[2] For Dennett's attempt to blunt this point, cf. Dennett (1981).

[3] See also Stich (1983, Ch. 5). Dennett's view is often described as of a piece with Davidson's. But this is clearly mistaken. Davidson makes no use of the notion of an ideally rational system. Like me, he insists that a person must be cognitively *similar* to ourselves if we are to succeed in understanding his speech and ascribing beliefs to him. In particular, he maintains that "if I am right in attributing a particular belief to you, then you must have a pattern of beliefs much like mine." (Davidson, 1979, p. 295). Davidson goes on to argue that most of these beliefs must be *true*. This is a view that Dennett holds as well. But as we shall see in the next section, Dennett's defense of this doctrine turns on evolutionary

considerations, while Davidson's does not. The least obscure argument Davidson offers for this conclusion goes like this: "There is nothing absurd in the idea of an omniscient interpreter". (Ibid.) To interpret us, this omniscient interpreter must share the bulk of our beliefs. And since *ex hypothesis* all of his beliefs are true, it follows that the bulk of ours must be true as well. End of argument. It should be pretty clear, however, that this argument simply begs the question. Granting the point about belief similarity being necessary for interpretation, it is an open question whether an omniscient interpreter could interpret our utterances as meaning something in his language. He could do so only if the bulk of our beliefs are true. And that is just what the argument was supposed to establish.

[4] Quoted in Sinsheimer (1971).

[5] We should note in passing that Cohen was not the first to introduce the competence/performance distinction into the debate about human rationality. Fodor (1981) has an extended and illuminating discussion of the possibility that "the postulates of . . . logic are mentally represented by the organism, and this mental representation contributes (in appropriate ways) to the causation of its beliefs" (p. 120). Since the internally represented logic would be only one among many interacting causes of belief and behavior, "the evidence for attributing a logic to an organism would not be that the organism believes whatever the logic entails. Rather, the appropriate form of argument is to show that the assumption that the organism internally represents the logic, when taken together with independently motivated theories of the character of the other interacting variables, yields the best explanation of the data about the organism's mental states and processes and/or the behaviors in which such processes eventuate". But if the facts turn out right, it would seem that the same sort of evidentiary considerations might also lead to the conclusion that the organism had internally represented a peculiar or normatively inappropriate "logic". This is not a possibility Fodor pursues, however, since he has been seduced by Dennett's argument from natural selection. Darwinian selection, he claims, "guarantees that organisms either know the elements of logic or become posthumous" (p. 121).

REFERENCES

Alcock, J.: 1979, *Animal Behavior: An Evolutionary Approach*, Sinauer Associates, Inc., Sunderland, MA.

Cohen, L. J.: 1981, 'Can Human Irrationality Be Experimentally Demonstrated?', *Behavioral and Brain Sciences* **4**, 3.

Davidson, D.: 1979, 'The Method of Truth in Metaphysics', in P. A. French, T. E. Uehling, Jr., and H. K. Wettstein (eds.), *Contemporary Perspectives in the Philosophy of Language*, University of Minnesota Press, Minneapolis.

Dennett, D.: 1978, *Brainstorms*, Bradford Books, Montgomery, VT.

Dennett, D.: 1981, 'Making Sense of Ourselves', *Philosophical Topics* **12**, 1.

Dennett, D.: 1981a,'Three Kinds of Intentional Psychology', in R. Healey (ed.), *Reduction, Time and Reality*, Cambridge Univ. Press, Cambridge.

Fodor, J.: 1981, 'Three Cheers for Propositional Attitudes', in *Representations: Philosophical Essays on the Foundations of Cognitive Science*, MIT Press and Bradford Books, Cambridge, MA.

Garcia, J., B. K. McGowan, and K. F. Green: 1972, 'Biological Constraints on

Conditioning', in A. H. Black and W. F. Prokasy (eds.), *Classical Conditioning II: Current Research and Theory*, Appleton-Century-Crofts, New York.

Griggs, R. A. and J. R. Cox: forthcoming 'The Elusive Thematic-Materials Effect in Wason's Selection Task', in *British Journal of Psychology*.

Johnson-Laird, P. N., P. Legrenzi, and M. Sonino Legrenzi: 1972, 'Reasoning and a Sense of Reality', *British Journal of Psychology* **63**.

Johnson-Laird, P. N. and P. C. Wason: 1970, 'A Theoretical Analysis of Insight Into a Reasoning Task' and 'Postscript – 1977', in P. N. Johnson-Laird and P. C. Wason (eds.), *Thinking*, Cambridge University Press, Cambridge.

Nisbett, E. W. and T. D. Wilson: 1977, 'Telling More than We Can Know: Verbal Reports on Mental Processes', *Psychological Review* **84**.

Nisbett, R. E. and L. Ross: 1980, *Human Inference: Strategies and Shortcomings of Social Judgement*, Prentice-Hall Inc., Englewood Cliffs, NJ.

Ross, L., M. R. Lepper, and M. Hubbard: 1975, 'Perseverance in Self Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm', *Journal of Personality and Social Psychology* **32**.

Sinsheimer, R. L.: 1971, 'The Brain of Pooh: An Essay on the Limits of Mind', *Science* **59**.

Slovic, P., B. Fischoff, and S. Lichtenstein: 1977, 'Behavioral Decision Theory', *Annual Review of Psychology* **28**.

Slovic, P. and B. Fischoff: 1978, 'How Safe is Safe Enough?' in L. Gould and C. A. Walker (eds.), *The Management of Nuclear Wastes*, Yale University Press, New Haven. Reprinted in J. Dowie and P. Lefrere (eds.), *Risk and Chance*, Milton Keynes, Open University Press.

Stich, S. P.: 1980, 'Headaches', *Philosophical Books* **21**, 2.

Stich, S. P.: 1981a, 'Dennett on Intentional Systems', *Philosophical Topics* **12**, 1.

Stich, S. P.: 1981b, 'On the Ascription of Content', in A. Woodfield (ed.), *Thought and Object*, Oxford University Press, Oxford.

Stich, S. P. and R. E. Nisbett: 1980, 'Justification and the Psychology of Human Reasoning', *Philosophy of Science*, **47**.

Stich, S. P.: 1983, *From Folk Psychology to Cognitive Science*, MIT Press, Cambridge, MA.

Tversky, A. and D. Kahneman: 1982, 'Judgments of and by Representativeness', in D. Kahneman, P. Slovic and A. Tversky (eds.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.

Wason, P. C.: 1977 'Self-Contradiction', in P. N. Johnson-Laird and P. C. Wason (eds.), *Thinking*, Cambridge University Press, Cambridge.

Wason, P. C. and P. N. Johnson-Laird: 1972, *The Psychology of Reasoning: Structure and Content*, B. T. Batsford, London.

Dept. of Philosophy
University of Maryland at College Park
College Park, MD 20742
U.S.A.