Philosophy and Phenomenological Research
Vol. LVII, No. 3, September 1997

Can We Be Justified in Believing That Humans Are Irrational?

EDWARD STEIN
Yale University

In this paper, the author considers an argument against the thesis that humans are irrational in the sense that we reason according to principles that differ from those we ought to follow. The argument begins by noting that if humans are irrational, we should not trust the results of our reasoning processes. If we are justified in believing that humans are irrational, then, since this belief results from a reasoning process, we should not accept this belief. The claim that humans are irrational is, thus, self-undermining. The author shows that this argument—and others like it—fails for several interesting reasons. In fact, there is nothing self-undermining about the claim that humans are irrational; empirical research to establish this claim does not face the sorts of a priori problems that some philosophers and psychologists have claimed it does

1.

The thesis that humans are irrational has become commonplace since Freud. In the past few decades, cognitive psychologists have performed a series of experiments that are supposed to prove that humans are irrational in a quite specific sense, namely that we reason according to principles that differ from those we ought to follow.[1] Many people see these experiments as having established that humans are irrational. Others have criticized these experiments on a variety of grounds. Some have criticized them on a piecemeal basis, pointing to particular flaws in their experimental methods.[2] Others have

[1] See, for example, Amos Tversky and Daniel Kahneman, "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," *Psychological Review* 90 (October 1983), 293–315; various articles in Daniel Kahneman, Paul Slovic and Amos Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases* (Cambridge: Cambridge University Press, 1982); Peter Wason, "Reasoning," in *New Horizons in Psychology*, Brian Foss, ed. (Middlesex, England: Penguin, 1966), 135–51; and various experiments discussed in Peter Wason and Philip Johnson-Laird, *Psychology of Reasoning: Structure and Content* (Cambridge: Harvard University Press, 1972).

[2] With regards to Kahneman and Tversky's work, see, for example, Gerd Gigerenzer, "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases,'" *European Review of Social Psychology* 2 (1991), 83-115. With regards to Wason and Johnson-Laird's work, see, for example, Leda Cosmides, "The Logic of Selection: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task," *Cognition* 31 (1989), 187–276.

argued that these experiments are based on mistaken assumptions about which principles of reasoning we ought to follow.[3] And still others, philosophers in particular, have made conceptual arguments that humans are rational and, thus, that these experiments should be interpreted in some other way than as showing that humans are irrational.[4]

In this paper, I am concerned with a particular conceptual argument and related arguments that make similar assumptions. The particular argument, simply put, says that the claim that humans are irrational undermines itself. According to this argument, we cannot be justified in believing that humans are irrational, because, if we were justified in believing this, then we should not trust the results of our reasoning processes, and, hence, we are not justified in believing that humans are irrational, since this conclusion is based on our reasoning process.[5] This argument is initially plausible and seems to have some force. In what follows, I examine this argument and its associated view about the relevance of empirical evidence to assessing human rationality. I also show that other arguments for the rationality thesis are interestingly similar to this argument.

My discussion proceeds as follows. In section 2, I clarify what it means to claim that humans are irrational. In section 3, I look at one example of the sort of empirical evidence that is supposed to support this claim. In section 4, I turn to the argument that the claim that humans are irrational is self-undermining. Having elaborated this argument, in section 5, I show what is wrong with it. In doing so, I further clarify the claim that humans are irra-

---

3 Gigerenzer, "How to Make Cognitive Illusions Disappear"; and L. Jonathan Cohen, "Can Human Irrationality Be Experimentally Demonstrated?," *Behavioral and Brain Sciences* 4 (1981), 317–70.

4 Three such arguments have to do with the principle of charity,—see the works of Daniel Dennett, especially *The Intentional Stance* (Cambridge: MIT Press, 1987); the works of Donald Davidson, especially *Inquires into Truth and Interpretation* (Oxford: Oxford University Press, 1984); and Elliott Sober, "Psychologism," *Journal of Social Behavior* 8 (1978), 165–91—the theory of reflective equilibrium,—see Cohen, "Can Irrationality Be Experimentally Demonstrated?"; and Edward Stein, "Rationality and Reflective Equilibrium," *Synthese* 99 (1994), 137-172—and the fact that humans are the result of evolution—see, for example, Daniel Dennett, "Making Sense of Ourselves," in *Intentional Stance*, 83-101; Karl Popper, "Evolutionary Epistemology," in *Evolutionary Theory: Paths into the Future*, Jeffrey Pollard, ed. (London: Wiley and Sons, 1984), 239–56; and Elliott Sober, "Evolution of Rationality," *Synthese* 46 (1981), 95-120. For discussions of all of these arguments see Stich, *Fragmentation of Reason* (Cambridge: The MIT Press, 1990); and Edward Stein, *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science* (Oxford: Oxford University Press, 1996).

5 John Macnamara, *A Border Dispute* (Cambridge: MIT Press, 1986), 184, explicitly makes this argument. Other friends of the claim that humans are rational seem to implicitly accept this argument or other arguments that make similar assumptions. See, for example, Cohen, "Can Human Irrationality Be Experimentally Demonstrated?"; and Sober, "Psychologism". Shades of this argument can also be found in arguments for the thesis that humans are rational that are related to the principle of charity; see note 4 above for relevant references.

tional and provide some guidelines for performing experiments to support it. In section 6, I consider and criticize two other arguments against the claim that humans are irrational; I show that these arguments have failings similar to the argument I articulated in section 4. I conclude in section 7.

2.

Before looking at the evidence for the thesis that humans are irrational (*the irrationality thesis*), I need to say more precisely what it would be for humans to be (or to fail to be) rational. Everyone agrees that humans frequently make mistakes in reasoning. Give a person a lot of alcohol, deprive her of sleep, or make her nervous, and her reasoning will be adversely affected. Those who think humans are rational and those who think they are not agree about situations like these. What, then, do they disagree about? At issue is the nature of our underlying ability to reason: is our underlying ability to reason—our *competence* for reasoning—structured in such a way that we will, when conditions are right (that is, when we are not drunk, not overtired, etc.), reason in the way that we ought to reason? People who think that humans are rational say that our *reasoning competence* enables us to reason as we ought to, while people who think humans are irrational say that our reasoning competence is such that we do not reason as we ought to.

A couple of these notions need to be explained further. When I talk about reasoning competence, I am talking about the human capacity for reasoning. I can have the capacity to do something (for example, ride a bike or apply *modus ponens*) and yet not display that capacity on a particular occasion (for example, because I am tired or drunk). The idea of a reasoning competence[6] is supposed to be analogous to the notion of linguistic competence.[7] A person's linguistic competence is her underlying ability to understand and utter grammatical sentences. Roughly, linguistic competence is the cognitive program that is involved in the learning and use of language. Reasoning competence then is our underlying ability to reason; it is the cognitive program involved in reasoning.[8] That humans are rational means that our reasoning competence embodies the principles of reasoning we ought to follow. I call these principles the *normative principles of reasoning*. They include principles stemming from logic, probability theory and the like. An example of a normative principle of reasoning is the *and-elimination principle*, which says that if you be-

---

6 Macnamara, *A Border Dispute* , uses the term 'mental logic', Stich, *Fragmentation of Reason* , uses the term 'psycho-logic', and Cohen, "Can Human Irrationality Be Experimentally Demonstrated?", uses the term 'cognitive competence', for roughly the same concept that I prefer to call 'reasoning competence'.

7 The notion of linguistic competence comes from so-called generative linguistics. See Noam Chomsky, *Rules and Representations* (New York: Columbia University Press, 1980); and Chomsky, *Knowledge of Language* (New York: Praeger, 1986).

8 For an extensive discussion of the notion of reasoning competence and the analogies and disanalogies with linguistic competence, see Stein, *Without Good Reason*, Chapter Two.

lieve the conjunctive statement **A and B**, you should believe both the statement **A** and the statement **B**. What I call the *rationality thesis* says that human reasoning competence embodies the normative principles of reasoning, while the irrationality thesis says that human reasoning competence embodies principles of reasoning that diverge from the norms. The argument that is the main concern of this paper, and which I take up in section 4, is supposed to show that the irrationality thesis is self-undermining. For now, I turn to the empirical evidence that is supposed to count in favor of the irrationality thesis.

3.

Consider the following description:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Given this description, rank the following statements in terms of their likelihood:

(1) Linda is active in the feminist movement.

(2) Linda is a bank teller.

(3) Linda is a bank teller and is active in the feminist movement.

This task was part of a series of experiments performed by psychologists Amos Tversky and Daniel Kahneman.[9] The experiments show that most people rate (1) the most likely, (3) the next most likely, and (2) the least likely. This, however, violates an important principle that is based on a rule of probability theory, *the conjunction principle*, which says that you should not attach a lesser degree of probability to an event **A** than you do to both **A** and a (distinct) event **B** occurring. In the case of Linda, you should not attach a lesser degree of probability to its being the case that Linda is a bank teller than you do to its being the case that Linda is a bank teller and a feminist. Every feminist bank teller is necessarily a bank teller but not all bank tellers are feminist. Linda could, despite her past, be a non-feminist bank teller, but nothing could make her a feminist bank teller who is not a bank teller. For this reason, if you thought that (3) is more likely than (2), you made a mistake. If you did, however, you need not blame your lapse on lack of sleep or your present mood; if Kahneman and Tversky are right, you were reasoning

[9] Tversky and Kahneman, "Extensional Versus Intuitive Reasoning."

in accordance with the principles embodied in human reasoning competence but these principles diverge from the normative principles of reasoning. In other words, you were just reasoning as humans do, namely, in an irrational manner.

At first glance, it might seem that there are more plausible ways of interpreting Kahneman and Tversky's data. Rather than seeing the experiment as showing that we lack the conjunction principle in our reasoning competence, why not say that subjects are simply misinterpreting the experiment? For example, subjects might be reading 'Linda is a bank teller.' as meaning that Linda is a bank teller, *but not* a feminist. If subjects are reading 'Linda is a bank teller.' in this way, it might explain why they think this statement is less likely to be true than 'Linda is a bank teller and a feminist.'; 'Linda is a bank teller *but not* a feminist.' would not be true in any of the same instances as 'Linda is a bank teller *and* a feminist.' whereas 'Linda is a bank teller.' would be true in some of the same instances as 'Linda is a bank teller and a feminist.' To test this possible explanation, subjects were presented with the following two statements about Linda and were asked to rank their probability:

(2′) Linda is a bank teller whether or not she is active in the feminist movement.

(3) Linda is a bank teller and is active in the feminist movement.

A majority of subjects rated (3) more probable than (2′). Subjects rated a conjunction as more probable than one of its conjuncts even though it was made clear that (2′) is in fact logically equivalent to a conjunct of (3); subjects neglect the conjunction principle even when they have been explicitly told that 'Linda is a bank teller.' does *not* mean that Linda is a *non-feminist* bank teller.[10] This suggests that subjects are not misinterpreting the task before them, but that they *are* in fact violating the conjunction principle and, thus, it seems that some other principle guides their reasoning.

There is a jump from the failure to use the conjunction principle to the absence of the conjunction principle in human reasoning competence. Perhaps subjects have the conjunction principle in their reasoning competence and know that it is the right principle to apply in the Linda case, but somehow fail to make the correct probability judgment in spite of this. One way to flesh out this suggestion is that subjects correctly interpret the task and call up the right principle, but they misapply it and thereby give the wrong answer. This suggestion has also been tested by Tversky and Kahneman. After being presented with the description of Linda and alternatives (2) and (3)—that Linda is a bank teller and that she is a feminist bank teller, respec-

[10] Ibid., 299.

tively—subjects were asked to indicate which of the two arguments for (2) and (3), respectively, they found the most convincing:

(A1) Linda is more likely to be a bank teller than she is to be a feminist bank teller because every feminist bank teller is a bank teller, but some women bank tellers are not feminists and Linda could be one of them.

(A2) Linda is more likely to be a feminist bank teller than she is likely to be a bank teller because she resembles an active feminist more than she resembles a bank teller.[11]

A majority of the subjects chose argument (A2) that advocates violating the conjunction principle. These results suggest that subjects apply some other principle even when the conjunction principle is spelled out for them.

Perhaps, as some have argued, there are problems with Tversky and Kahneman's results.[12] There is, however, evidence from a wide range of other experiments that suggests humans do not reason in accordance with the normative principles.[13] Such evidence seems to provide strong empirical evidence for the irrationality thesis. The argument that I develop in the section that follows, however, is suppose to show that, even in the face of such evidence, we are not justified in believing that humans are irrational.

4.

The initial insight behind the argument against the irrationality thesis comes from the observation that to inquire whether humans are rational, we must use the very reasoning capacities that we are attempting to assess. If we are irrational, then our inquiry into human rationality (as well as our other scientific and philosophical inquiries) is suspect because engaging in this inquiry requires that we reason, but, as reasoners, we make systematic errors. As a result, if the irrationality thesis is true, we cannot know that it is true; insofar as evidence leads us to believe the irrationality thesis, this evidence suggests that we cannot be justified in believing the irrationality thesis. According to this argument, the irrationality thesis is (roughly) self-refuting: if the irrationality thesis is true, then the evidence that is supposed to establish its truth is either false or it does not in fact establish its truth. The *rationality* thesis does not face a parallel problem. If we are rational, then it is at least possible that our inquiry into human rationality will not be suspect; if,

---

11 Ibid., 299.

12 See, for example, Gigerenzer, "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases.'"

13 See, for example, various articles in Kahneman, Slovic and Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases* and various experiments discussed in Wason and Johnson-Laird, *Psychology of Reasoning*.

through reasoning, we come to the conclusion that humans are rational, our conclusion is perfectly consistent with the reliability of the methods used to discover it. The rationality thesis is thus not self-refuting—it is compatible with its own truth. If the argument shows that the irrationality thesis is self-refuting and the rationality thesis is not, then this argument amounts to a *reductio ad absurdum* argument for the rationality thesis. If the irrationality thesis is self-refuting, then it leads to a contradiction, and, thus, the rationality thesis—the negation of the irrationality thesis—is true. The question then is whether in fact the irrationality thesis is self-refuting. Before I can answer this question, I need to distinguish among different senses of what self-refutation is.

A. Self-Refutation

A statement is self-refuting if it contradicts itself. There are three different senses in which a statement can be self-refuting.[14] First, a statement can be *pragmatically* self-refuting. Consider the sentence 'George cannot speak.' spoken by George. The sentence may well be true, for all we know, and it may well be true that it can be spoken by George; it cannot, however, be true *if* George speaks it. This is a paradigmatic case of pragmatic self-refutation: the sentence is only refuted if it is asserted in a certain way. Strictly speaking, however, 'George cannot speak.' is not self-refuting, but the action of uttering it is (the sentence is contradicted by George's *act* of speaking it); that is the fingerprint of pragmatic self-refutation.

Second, a statement can be *absolutely* self-refuting. Consider the sentence 'I know that I know nothing.' The sentence cannot be true, because the sentence asserts that I know something while at the same time asserting that I know nothing. The sentence is a contradiction; under no circumstances, under no way of putting forth this statement, can it be consistently asserted.

Third, a statement can be *operationally* self-refuting. Consider the sentence 'I believe that I have no beliefs.' If we assume that my asserting 'I believe **X**.' entails that I have at least one belief, the sentence 'I believe that I have no beliefs.' is self-refuting because it denies that I have any beliefs while at the same time, it entails that I have at least one belief. But unlike statements that are absolutely self-refuting, operationally self-refuting statements, although they cannot be consistently asserted in any fashion, could well be true; I could believe that I have no beliefs (this would be true, for example, if I was an eliminativist about beliefs[15]). Unlike pragmatically self-refuting

---

[14] John Mackie, "Self-Refutation—A Formal Analysis," *The Philosophical Quarterly* 14 (July 1964), 193–203, reprinted in *Logic and Knowledge: Selected Papers of J. L. Mackie*, Joan Mackie and Penelope Mackie, eds. (Oxford: Oxford University Press, 1985), 54–67.

[15] Eliminativism is the view that our commonsense psychological categories and the theory that underlies them are radically mistaken. According to this theory, there are, in fact, no

statements, in which the way that the sentence is asserted conflicts with the statement—George cannot say that he cannot speak, but he could use a sign language to communicate it—with operationally self-refuting statements, there is no way that the sentence can be consistently put forth—even if I use a sign language, I cannot consistently assert that I believe that I have no beliefs. A pragmatically self-refuting statement, although it is self-refuting if asserted in a certain way by a person (or set of people), can be consistently asserted in *some* way by those who could not assert it in the first way. A statement is operationally self-refuting if it is not absolutely self-refuting and it cannot be asserted in *any* way by a person (or set of people); a statement that is operationally self-refuting can, however, be true.

B. The Maximal Irrationality Thesis

Before I consider whether the irrationality thesis is self-refuting, I want to consider a thesis that is even more plausibly thought to be self-refuting, namely the thesis that *every* inference humans make violates the normative principles of reasoning—what I call the *maximal irrationality thesis* (the *maximal thesis*, for short). If the maximal thesis is true, then so is the irrationality thesis. The reverse is not true: even if the irrationality thesis is true, the maximal thesis could still be false. If the maximal thesis is not self-refuting, then the irrationality thesis cannot be, because the irrationality thesis is entailed by the maximal thesis. If, however, the maximal thesis is self-refuting, this does not entail that the irrationality thesis is self-refuting (though it is consistent with its being so).

Would any empirical evidence justify the belief that *every* inference humans make is irrational? Can I assert that *every* inference humans make is irrational? The idea is that the answer to these questions is no; the maximal thesis is self-refuting because I must make some inferences in order to arrive at any general beliefs about humans, but according to the maximal thesis, all inferences humans make violate the norms. In what way, if any, is the maximal thesis self-refuting? The claim that humans are maximally irrational is not pragmatically self-refuting: the mode of presentation does not affect the consistency of such a statement. It does not affect the consistency of the statement whether I speak or write it. Further, the maximal thesis is not absolutely self-refuting; it could be the case that every inference humans make is irrational—this statement is not self-contradictory. It seems, then, that the maximal thesis might be operationally self-refuting. Consider the version of the maximal thesis that makes the claim "I have concluded that every inference humans make violates the norms of reasoning." My asserting this ver-

---

such things as beliefs. See Paul Churchland, *Scientific Realism and the Plasticity of Mind* (Cambridge: Cambridge University Press, 1979); and Stephen Stich, *From Folk Psychology to Cognitive Science: The Case Against Belief* (Cambridge: MIT Press, 1983).

sion of the maximal thesis suggests that I have reason to believe it, but if the statement is true, I cannot have any reason to believe it (or anything else for that matter) because, according to the statement itself, every belief that I have as the result of an inference is unjustified. The upshot is that although the maximal thesis could not be supported by empirical evidence or rationally believed by any human being, it still could be true.

There are two things to note here. First, the problem with the maximal thesis is different from it being absolutely self-refuting. If a statement is absolutely self-refuting, then its negation must be true. The same is not true for a statement that is operationally self-refuting; a statement that is operationally self-refuting is epistemologically inaccessible for those that it applies to. If the maximal thesis is operationally self-refuting, then the maximal thesis is epistemologically inaccessible to humans. Second, the problem with the maximal thesis is not even that it is operationally self-refuting. A statement is operationally self-refuting if it cannot be asserted in any way by a person or set of people. The maximal thesis can be consistently asserted; its problem has to do with justification rather than truth. My example of an operationally self-refuting statement was 'I believe that I have no beliefs.' This sentence has a problem with truth: if it is true that I believe that I have no beliefs, then it is false that I have no beliefs. The maximal thesis is more like the sentence 'I believe that none of my beliefs are justified.' This sentence has a problem with justification, not truth: if it is true that I believe that none of my beliefs are justified, I can still consistently believe that none of my beliefs are justified; I cannot, however, consistently believe that I am *justified* in believing that no beliefs are justified. With respect to the maximal thesis, I can consistently believe that every inference humans make violates the norms, but I cannot consistently believe that I am justified in this belief. This still allows that the maximal thesis is epistemologically inaccessible to humans because it is self-undermining, but it is not straightforwardly self-refuting, not even operationally self-refuting.

C. Epistemological Inaccessibility

The epistemological situation of the maximal irrationality thesis is similar to a standard account of a person's epistemological situation with respect to other minds. The problem of other minds has to do with whether a person can know that there are other minds in the world. Of course, I can see there are other bodies, but my only evidence that there are other minds has to do with the behavior of these other bodies. The observed behavior of other bodies is perfectly consistent with it being the case that these other bodies have no minds and instead are robots or zombies. It is possible that there are other minds, but it is also possible that there are *not*. This question, however, may be unanswerable. There may well be no evidence I can have access to that

will bear on this question. The sort of evidence that would be relevant (for example, that some other body has conscious experience or feels pain) is epistemologically inaccessible. This shows that even if the maximal irrationality thesis is epistemologically inaccessible, it is in good company. We might be able to be as confident in asserting that humans are maximally irrational as I am that there are other minds besides my own.

The irrationality thesis cannot be worse off in terms of self-refutation then the maximal irrationality thesis because the irrationality thesis is contained in the maximal irrationality thesis. Even if the maximal irrationality thesis were operationally self-refuting, the strongest claim of self-refutation that could be made against the irrationality thesis is that it is also operationally self-refuting. The strongest claim the argument under consideration could make against the irrationality thesis is that this thesis is epistemologically inaccessible. In fact, the argument under consideration is even weaker than this. The maximal thesis does not have the problem with truth that is the earmark of operational self-refutation. It has, rather, a similar problem with respect to justification. Further, even if the argument establishes that the irrationality thesis is epistemologically inaccessible, this would not in itself—as the analogy with the problem of other minds suggests—constitute an argument for the rationality thesis. It would, however, somewhat undermine the main argument for the irrationality thesis, because it would undermine the empirical evidence that is supposed to support it. The rationality thesis is not epistemologically inaccessible in the way that the irrationality thesis is. This asymmetry with respect to epistemological accessibility suggests that the rationality thesis is in better shape than the irrationality thesis. If the argument is right, we can never have evidence for believing the irrationality thesis, while there is at least the possibility that some evidence will support the rationality thesis. In the next section, I turn to evaluating this argument.

5.

Consider the following version of the argument that we cannot be justified in believing the irrationality thesis:

(1) We are justified in believing that humans are irrational in light of empirical evidence. [assumption]

(2) If humans are irrational, then we (as humans) cannot be justified in believing anything that relies on our reasoning process.

(3) We cannot be justified in believing that humans are irrational on the basis of empirical evidence. [from 1 and 2]

(4) Since (1) and (3) contradict each other, we should reject (1); therefore, we have no empirical evidence that humans are irrational.

The crucial step in the argument is (2), which initially seems reasonable. The irrationality thesis says that humans systematically violate the norms of reasoning in the sense that human reasoning competence diverges from the norms of reasoning. The idea behind (2) is that if our reasoning competence diverges from the norms then we can never be justified in any of our beliefs. Is this right?

A. A Specific Example

Unlike the maximal irrationality thesis, the irrationality thesis does not claim that *every* inference humans make is irrational; it only claims that *some* of the principles we have in our reasoning competence diverge from the normative ones. For example, the irrationality thesis claims that humans do not, in virtue of our reasoning competence, follow the conjunction principle in certain situations in which we should follow it. Is the claim that humans do not follow the conjunction principle self-refuting? Consider the argument rewritten for the specific case of the Linda experiment:

(C1) We should believe that humans lack the conjunction principle in their reasoning competence on the evidence of the Linda experiment. [assumption]

(C2) If humans lack the conjunction principle in our reasoning competence, then humans are not justified in believing any conclusion that requires humans to reason rationally.

(C3) Humans are not justified in believing that humans lack the conjunction principle in our reasoning competence on the basis of the Linda experiment. [from C1 and C2]

(C4) Since (C1) and (C3) contradict each other, we should reject (C1); therefore, the Linda experiment does not count as evidence that humans lack the conjunction principle in their reasoning competence.

What do we make of this argument, in particular, what do we make of (C2)? (C2) basically says that if humans lack the conjunction principle, then human reasoning cannot prove anything and cannot be relied upon. The idea is that, given the ubiquity of the conjunction principle in human reasoning, if humans do not reason in accordance with the conjunction principle, then the results of human reasoning cannot be trusted. There are at least two problems with (C2) and this rationale for believing it.

Although there are many situations in which the conjunction principle might be invoked, it is not the case that every time I acquire a new belief I need to invoke the conjunction principle. Lacking the conjunction principle does not thereby undermine all of my reasoning processes; at best it undermines reasoning that requires the invocation of the conjunction principle. This suggests that (C2) ought to be rewritten as follows:

(C2′) If humans lack the conjunction principle in our reasoning competence, then humans are not justified in believing any conclusion of a reasoning process that demands the invocation of the conjunction principle.

(C2′) seems more plausible than (C2). There is still, however, a problem with it. The conjunction principle says that you should not attach a lesser degree of probability to an event **A** than you do to both the event **A** and the (distinct) event **B**. The Linda experiment does not show that people *always* fail to attach a lesser degree of probability to an event **A** than they do to both **A** and **B**, only that they *sometimes* do, in particular, that they typically do in situations like the one involving Linda. Although humans may lack the conjunction principle in our reasoning competence, we may have a principle in its place that, *in certain contexts*, results in the same inferential behavior as the conjunction principle does. In such contexts, it seems that we *are* justified in believing the conclusions of such inferences because, in these contexts, inferences in accordance with such principles match the inferences in accordance with the norms. This suggests a further modification to (C2). Consider:

(C2′′) If humans lack the conjunction principle in our reasoning competence and instead have some other principle **P**, then humans are not justified in believing any conclusion that is based on **P** rather than the conjunction principle unless **P** and the conjunction principle result in the same conclusion in the particular context.

(C2′′) seems true, but it is not at all clear that (C2′′) will fill (C2)'s spot in the argument for (C4). (C1) and (C2) were supposed to entail the negation of (C1). Does (C2′′), when conjoined with (C1), entail (C3), the negation of (C1)? This depends on whether the process of forming the belief that humans lack the conjunction principle requires the use of the conjunction principle in a context in which the principle that we have in our reasoning competence in place of the conjunction principle—**P**—diverges from the conjunction principle. If the process of forming the belief "Humans lack the conjunction principle." requires the use of the conjunction principle in contexts in which **P** diverges from it, then it seems that (C1) and (C2′′) will entail (C3). It is not at all clear, however, that this is the case.

B. Does Irrational Mean Always *Irrational?*

Rather than dwell on the particular version of the argument that involves the conjunction principle, I want to return to the general version of the argument with these two objections in hand. Recall premise (2) of the argument against the rationality thesis:

> (2) If humans are irrational, then we cannot be justified in believing anything that relies on our reasoning process.

What precisely does (2) mean? The irrationality thesis says that humans lack the normative principles of reasoning in our reasoning competence. This entails that there are many contexts in which we will reason in accordance with our reasoning competence but fail to reason in accordance with the norms. For (2) to be true, *every* instance of human reasoning must fall into one of these contexts, that is, every instance of human reasoning must involve a principle of reasoning that diverges from the norms in a context in which that principle produces an inference that is different from what the relevant norm would produce. (2) says that if the irrationality thesis is true, then humans cannot be justified in believing anything. The idea is that if we are irrational, any inference we make will be tainted by our irrationality. It seems highly unlikely that this is true.

My discussion of the specific version of the argument against the irrationality thesis (discussed in section A above) suggests two objections to (2). With respect to the specific version of the argument, the first objection was that the conjunction principle is not invoked every time we engage in reasoning. Generalizing this objection, one objection to (2) is that the non-normative principles that we have in our reasoning competence are not invoked every time we reason. The reasoning experiments do not provide evidence for the view that *none* of the principles in our reasoning competence match the norms and the irrationality thesis is not committed to this view. (2) seems in trouble because we can be justified in believing the results of a reasoning process that relies on those normative principles that we do in fact have in our reasoning competence.

As an example, consider how a reductio-style argument relating to vision might be developed on the model of this argument. An experimenter studying the bent-stick-in-water illusion must, of course, use vision to perform her experiments. It would be ridiculous to claim that the experimenter's conclusion that humans are subject to visual illusions calls *all* visual data into question. Visual illusions occur only in certain contexts. The same sort of claim is true with respect to the reasoning experiments; even if we are irrational, it does not follow that all human reasoning is bad reasoning.

The second objection to the particular version of the argument against the irrationality thesis was that even though we are reasoning in accordance with

some principle other than the conjunction principle we do not *always* fail to make inferences that accord with the conjunction principle. Generalizing this objection, it is possible for me to engage in the same reasoning behavior I would if I had some normative principle **N** in my reasoning competence even though I have some non-normative principle **P** in my reasoning competence. In other words, while I fail to reason in accordance with **N** in certain contexts, there may well be other contexts in which reasoning in accordance with **P** will produce the same reasoning behavior as reasoning in accordance with **N**.

None of the experiments that are supposed to provide evidence for the irrationality thesis claim that we *never* reason in the manner that we would if we had the normative principles of reasoning in our reasoning competence. Rather, they allow that there may be some contexts in which we reason as we would *if* we had the normative principles in our reasoning competence while insisting that we in fact do not have the normative principles there. Given this, (2)—which says that if the irrationality thesis is true, humans cannot be justified in believing anything that results from our reasoning—seems in deep trouble; we can be justified in believing the results of our reasoning so long as we are relying on principles of reasoning in contexts in which these principles produce reasoning behavior that accords with the normative principles of reasoning.

C. The "Careful" Cognitive Scientist

Friends of the argument against the irrationality thesis might attempt to revise (2) to respond to these two objections. Consider:

> (2′) If humans are irrational, then we (as humans) cannot be justified in believing any conclusion based on a non-normative principle of reasoning unless the reasoning based on that principle was done in a context in which reasoning based on it accords with reasoning based on the appropriate normative principle of reasoning.

This premise is immune to the two objections raised against (2) above. The question, however, is whether (2′) can take the place of (2) in the argument against the irrationality thesis. For the argument with (2′) taking the place of (2) to produce the necessary contradiction, it must be required that, in order to come to believe the irrationality thesis based on empirical evidence, we must rely on non-normative principles of reasoning in contexts in which they diverge from the reasoning behavior that would result from the norms. The initial version of the argument obtained a contradiction from the premise that we have empirical evidence for the irrationality thesis by arguing that the truth of the irrationality thesis entails that no conclusion based on human reasoning can be justified. I have tried to show that, under closer examination, the irra-

tionality thesis does not entail that human reasoning is so impotent. Conclusions made on the basis of human reasoning can be justified even if the irrationality thesis is true so long as they are based on either normative principles of reasoning or on non-normative principles in contexts in which inferential behavior based upon such principles matches the inferential behavior associated with the norms. To establish a contradiction between the fact that humans are irrational and the fact that human reasoning is required to come to know this fact, friends of this argument need to show that every chain of argument leading to the conclusion that humans are irrational is based on a non-normative principle of reasoning in a context in which it diverges from the inferential behavior of the associated norms.

This seems a rather difficult task to accomplish. There are many chains of reasoning that lead to a particular conclusion. Even if one could show that a particular chain of reasoning to some conclusion requires a norm that humans lack, this does not show that there is *no* chain of reasoning to that conclusion which involves inferences based on principles we have in our reasoning competence in contexts in which they match the norms. For the argument to work, it needs to be shown that, regardless of the evidence that might be uncovered, *every* chain of reasoning to the conclusion that the irrationality thesis is true requires an inference that can only be based on a principle of reasoning that human reasoning competence lacks.

It seems to me unlikely that all chains of reasoning that lead to the irrationality thesis, regardless of the evidence they invoke, require inferences that must be based on principles that humans lack in our reasoning competence. It seems possible that there could be a "careful" cognitive scientist who avoided reasoning using principles in contexts in which they diverge from the normative principles of reasoning. Such a cognitive scientist might be able to come to the conclusion that humans are irrational without relying on any principle of reasoning in a context in which it diverges from the norms. If this is possible, then the truth of (2′) is compatible with humans being justified in believing in the irrationality thesis, and hence that the irrationality thesis is not self-undermining.

One might suspect, however, that there is something tricky going on here.[16] How could a cognitive scientist reliably know which principles of reasoning to avoid in each of various contexts? In order to conclude that some principle **P** produces reasoning behavior that, in some context, matches the reasoning behavior that would be produced by the relevant norm **N**, the cog-

[16] Most friends of the rationality thesis—for example, Cohen, "Can Human Irrationality Be Experimentally Demonstrated?"—would be suspicious of such a possibility. For a more specific statement of this suspicion, see Cohen, "Reply to Stein," *Synthese* 99 (1994), 173–76, especially 175–76. For a response to worries like Cohen's that differs somewhat from what follows here, see Edward Stein, "Cordoning Competence," *Synthese* 99 (1994), 177–79.

nitive scientist could not rely on any principles that produce inferences that diverge from those produced by the relevant norms. It seems that if she were careful, she would be able to do this. There is a problem, however: how would the process get started? In other words, in making her *first* conclusion that humans lack some normative principle in our reasoning competence, how could the careful cognitive scientist be sure that she is not *already* relying on some non-normative principle of reasoning that undermines her justification for believing this conclusion?

The cognitive scientist cannot be sure that she is not, from the start of her investigation, in some way basing her reasoning on a principle that diverges from the norms. She can, however, decrease the chances that she is doing this. The chances that she is, from the start, using a *non*-normative principle as a part of her reasoning method can be reduced if, once her investigation is underway, she checks back over her initial reasoning to see if any of the principles she used from the start are non-normative ones. This process of self-checking will not *guarantee* that the cognitive scientist is not using any non-normative principles. The cognitive scientist might be using a principle that she has not yet discovered is in our reasoning competence or she might be using a principle that is non-normative although she has not yet determined that it is. Despite this, the practice of self-checking for irrationality would dramatically reduce the chances that one is using non-normative principles of reasoning, and thereby further weaken the argument against the irrationality thesis.

Further, we might imagine that not only is the cognitive scientist *careful*, she is also *lucky*. Not only does this cognitive scientist self-check to make sure that she is not using some of the very principles that she has discovered are non-normative and in our reasoning competence, but, further, none of the principles in human reasoning competence that she is reasoning in accordance with as part of her investigation of human rationality in fact diverge from the norms in the contexts she is using them. For matters to work out this way, she has to be quite lucky. In spite of how much luck is involved, this situation is *possible* (although a person will have no way of knowing for sure whether she is this lucky). In fact, if enough people are engaging in this kind of research, the odds may be good that *some* cognitive scientist will be both careful and lucky and will, successfully, and without using any non-normative principles of reasoning, prove that humans are irrational. Given this possibility, (3) does not follow from (1) and (2′) and thus the charge that the irrationality thesis is self-undermining fails to stick.

The careful cognitive scientist example brings out an interesting point. Even if humans lack a particular normative principle in our reasoning competence, we might come to appreciate that this principle is a normative one and

even try to bring our reasoning into accordance with it.[17] This point is further supported by an analogy with linguistics. It follows from linguistic theory that there are *conceivable* (though non-human) languages that do not share some of the features all possible *human* languages share. A human child, brought up among beings (call them Martians) who spoke such a non-human language (call it Martianese), would not be able to acquire the language of her adoptive family with the same remarkable speed with which human children, raised by humans, normally acquire human languages.[18] This does not mean she would never be able to learn Martianese and communicate with Martians. The crucial difference for humans between learning Martianese and English is that we have a specialized capacity for learning languages like English but not for learning languages like Martianese. This does not, however, mean we could not learn Martianese—after all, we learn to do lots of things for which we have no specialized competence.

To take a specific example with respect to reasoning, it seems possible that, having determined that humans lack the conjunction principle, we might endeavor to teach people to recognize the various situations in which the conjunction principle should be invoked and, then, to apply the conjunction principle in these situations. People trained in this way might still violate the conjunction principle in unfamiliar contexts or when they are rushed or nervous, but they might reason in accordance with it more frequently than untrained people would. Something like this occurs in the case of visual illusions. We learn not to believe our eyes when the stick in the water looks bent. We think to ourselves, "I know the stick *looks* bent, but this is just one of those visual illusions; the stick is really straight." People trained to follow the conjunction principle might, when presented with the case of Linda, think, "I am tempted to say that Linda is more likely to be a bank teller and a feminist than she is to be a bank teller, but I recognize that this is one of those tricky situations; Linda is really more likely to be a bank teller." That this seems possible lends further support to the view that even if the irrationality thesis is right and we lack certain normative principles in our reasoning competence, it does not necessarily follow that some people will not be able to learn such principles and, when they are being careful and

---

17 For a discussion of actual attempts to teach people principles that are not part of their reasoning competence, see, for example, R. P. Larrick, J. N. Morgan, and R. E. Nisbett, "Teaching the Use of Cost-Benefit Reasoning in Everyday Life," in Richard Nisbett, ed., *Rules for Reasoning* (Hillsdale, NJ: Lawrence Erlbaum, 1993), 259–76; and R. E. Nisbett, G. T. Fong and D. R. Lehman, "Teaching Reasoning" in Nisbett, *Rules for Reasoning*, 297–314. For a more abstract discussion, see also, Stich, *The Fragmentation of Reason*, *passim*, especially Chapter Six.

18 Hilary Putnam, "The 'Innateness Hypothesis' and Explanatory Models in Linguistics," *Synthese* 17 (1967), 12–22, reprinted in Ned Block, *Readings in the Philosophy of Psychology*, volume two (Cambridge: Harvard University Press, 1981), 292–99, makes the stronger and mistaken claim that linguists are committed to the view that humans would be *unable* to learn Martianese (292, reprinted version).

reflective, that they will be able to follow them. For this further reason, the argument against the irrationality thesis fails.

6.

I have shown that there are two main problems with the argument against the irrationality thesis. First, at best, the argument can only establish that the irrationality thesis is epistemologically inaccessible, not that it is false and not that it is self-refuting. Second, the irrationality thesis does not entail that inferential behavior based on human reasoning competence will always diverge from the norms; given this, the irrationality thesis does not undermine itself because humans can be somewhat irrational without being irrational in such a way that makes human reasoning impotent. None of what I have shown so far establishes the irrationality thesis. In general, friends of the rationality thesis have two options for undermining the evidence that seems to support human irrationality. First, they might try to point to problems with the particular experiments. This approach has shown promise,[19] but it necessarily proceeds on piecemeal basis and thus fails to provide a reason for thinking there is something wrong with *every* experiment that claims to show that humans are irrational.[20] Second, they might try to undermine the evidence that is supposed to support the irrationality thesis by providing a conceptual argument against the irrationality thesis. The bulk of this paper has been devoted to considering one such argument, namely that the irrationality thesis is self-undermining. I have shown that this argument fails. There are, however, other conceptual arguments for the rationality thesis. Some of these arguments fail for reasons related to the reasons why this argument fails. I now consider two such arguments.

Some people have made the following observations about experiments like the one involving Linda: notice that experimenters have no trouble recognizing that the subjects in these experiments are making mistakes.[21] Further, even though we recognize that we might make the same mistakes subjects do, we also see that subjects are making mistakes. For example, even though you might have originally thought that Linda is more likely to be a bank teller and a feminist than she is to be a bank teller, you came to realize that this is a mistake upon reading my discussion of the conjunction principle. Even the subjects themselves may come to see their own mistakes. The idea that follows from observations of this kind is that our ability to recognize divergences from the norms suggests that these principles are in our reasoning competence; in other words, if we can recognize that a principle of

---

19 See note 3, above.

20 For further elaboration of this point, see Stein, *Without Good Reason*, Chapter Three.

21 See, for example, Macnamara, *A Border Dispute*.

reasoning is a norm, then the principle must be in our reasoning competence.[22]

Although this may seem a plausible line of thought, I have already suggested that there are problems with this idea. There may be contexts in which we can recognize the mistakes in reasoning that we make, but this fact does not entail that the mistakes that we make are not due to our reasoning competence. We may recognize that we think the stick in the water is bent when it is really straight, but this does not change the fact that we see it as bent. Even if we recognize that we make mistakes in reasoning, we do not thereby immunize ourselves against making them again. We might recognize that a principle of reasoning is one that we have been told is a norm without in fact using such a principle in most contexts or without having the principle in our reasoning competence. The present argument, just like the argument that the irrationality thesis is self-undermining, has a mistaken picture of reasoning competence. Both arguments assume that if humans can entertain a principle of reasoning, then this principle must be part of human reasoning competence. This does not seem to be the case at all.

Another conceptual argument against the irrationality thesis that shares some features with the arguments considered thus far is the *reflective equilibrium* argument. Reflective equilibrium is an epistemological theory that says a set of principles is justified when it is modified to fit with first-order intuitions about its domain of application.[23] According to the reflective equilibrium argument for the rationality thesis, both the normative principles of reasoning and the principles that characterize reasoning competence come from a process of reflective equilibrium with our intuitions about what counts as good reasoning as input. As such, the two sets of principles cannot diverge—humans must be rational; experiments that show the contrary must either be experimentally flawed or we must be misinterpreting their results.[24]

This argument has been criticized by saying that the normative principles of reasoning do not result from a process of reflective equilibrium.[25] The reflective equilibrium account of the norms has been defended by showing that critics of this account underestimate its resources.[26] Even if the reflective equilibrium account of the norms is right, I think the reflective equilibrium

---

22 Cohen, "Can Human Irrationality Be Experimentally Demonstrated?"; and Macnamara, *A Border Dispute*.

23 The concept of reflective equilibrium comes from Nelson Goodman, *Fact, Fiction and Forecast*, fourth edition (Cambridge: Harvard University Press, 1983), 63–64; it was taken up in John Rawls, *A Theory of Justice* (Cambridge: Harvard University Press, 1971), and baptized "reflective equilibrium" in John Rawls, "The Independence of Moral Theory," *Proceedings and Addresses of the American Philosophical Association* 48 (1974–75), 5–22.

24 Cohen, "Can Human Irrationality Be Experimentally Demonstrated?"

25 For example, Stich, *Fragmentation*, Chapter Four.

26 See Stein, "Rationality and Reflective Equilibrium."

argument for the rationality thesis still fails. Describing human reasoning competence might involve reflective equilibrium, but this does not guarantee that the principles that result from the two reflective equilibrium processes—the one to determine the norms and the other to characterize human reasoning competence—will be the same. There is, in fact, good reason for thinking the results of the two processes will diverge: the inputs to a reflective equilibrium process to determine our reasoning competence are different from the inputs to such a process to determine the norms of reasoning. According to the reflective equilibrium account, the normative principles of reasoning result from balancing our particular intuitions of what counts as good reasoning with general criteria of good reasoning, and, perhaps, with certain theoretical and philosophical considerations. In contrast, certain scientific evidence is relevant to developing an account of our reasoning competence. Empirical evidence thus is relevant to determining our reasoning competence but not to determining the norms. Given the different inputs to the two reflective equilibrium processes—one to determine the norms and one to determine reasoning competence—the principles that result from the two processes may differ, hence the reflective equilibrium argument fails to demonstrate that humans must be rational.

Some have argued that scientific evidence is relevant to determining what the norms are; they think that rationality should be naturalized.[27] If they are right, then it is possible that the input to the reflective equilibrium process to determine what the norms are and the reflective equilibrium process to determine what principles characterize human reasoning competence might be the same. Even if this were the case, different parts of the input would be weighted in different ways as part of the two reflective equilibrium processes. The goal of one process is to develop a descriptive psychological account of human reasoning competence and the goal of the other is to develop an account of the normative principles of reasoning. Even if the two reflective equilibrium processes get the same data as input, given their different goals, there is no reason to think the balancing process involved in developing a psychological theory would parallel the balancing involved in justification. Even if rationality is naturalized, reflective equilibrium does not produce a successful argument for the rationality thesis.[28]

If the reflective equilibrium argument against the irrationality thesis were right, it would entail that if we could recognize that a principle is a norm, then that principle must be in our reasoning competence. The other two arguments against the irrationality thesis that I have discussed make this same claim. In fact, such a claim seems intertwined with the reasons why all three of these conceptual arguments against the irrationality thesis fail. One of the

27 For a discussion, see Stein, *Without Good Reason*, 253–65 and 269–72.

28 Stein, "Rationality and Reflective Equilibrium."

problems that I discussed concerning the argument that the irrationality thesis undermines itself is also a problem for several other arguments against the irrationality thesis.[29]

VII.

My overall conclusion might be described as "mak[ing] the world safe for irrationality."[30] Although I have not proven that humans are irrational or even that human rationality is an empirical question, I have criticized certain arguments against the empirical evidence for human irrationality. I have argued that one can consistently believe that humans are irrational on the basis of empirical evidence concerning human reasoning; we may well have principles in our reasoning competence that diverge from the normative principles of reasoning. This does not, however, prevent us from recognizing that we are irrational, from being able to identify normative principles that are not in our reasoning competence, or from coming to learn to apply such normative principles in certain contexts. Failure to appreciate these points plagues various arguments for the view that humans must be rational.[31]

---

29 Not all conceptual arguments against the irrationality thesis involve the claim that our ability to recognize that a principle is a norm entails that the principle is in our reasoning competence. Neither do they fail for the reasons similar to those that plague the reflective equilibrium argument, the argument that the irrationality thesis is self-undermining, or the argument that humans must be rational because we can recognize instances of irrationality.

30 The phrase is borrowed from Stich, *Fragmentation*, 17.

31 I would like to thank Paul Bloom, John Gibbons, Peter Lipton, Roy Sorensen, Steve Stich and two anonymous referees for their helpful comments on earlier versions of this paper.