EDWARD STEIN

RATIONALITY AND REFLECTIVE EQUILIBRIUM

ABSTRACT. Cohen (1981) and others have made an interesting argument for the thesis that humans are rational: normative principles of reasoning and actual human reasoning ability cannot diverge because both are determined by the same process involving our intuitions about what constitutes good reasoning as a starting point. Perhaps the most sophisticated version of this argument sees reflective equilibrium as the process that determines both what the norms of reasoning are and what actual cognitive competence is. In this essay, I will evaluate both the general argument that humans are rational and the reflective equilibrium argument for the same thesis. While I find both accounts initially appealing, I will argue that neither successfully establishes that humans are rational.

1.

Aristotle's assertion that man is a rational animal seemed obviously true in his day. In contrast, today it is commonly accepted that humans are irrational. This commonly accepted observation about humans and their intellectual capacities has garnered scientific support in the last quarter century from some psychologists and cognitive scientists who have performed various experiments that are supposed to show that humans make systematic errors in reasoning and are therefore irrational.[1] In reaction to these experiments and the pronouncements which have been made based on them, various philosophers and others have defended the claim that humans are rational with a diverse set of interesting and plausible arguments.[2] In this essay, I focus on a particular type of argument for the claim that humans are rational. The conclusion of this argument is that the norms of reasoning and actual human reasoning ability have the same source, and, as such, must not diverge.

The claims that humans are irrational (what I call *the irrationality thesis*) and that humans are rational (what I call *the rationality thesis*) are based on a natural picture of what it is to be rational: to be rational is to follow the normative rules of reasoning. These norms include rules of logic, probability theory, and the like. For example, a person who systematically believes both *p* and *p implies q* but does not believe *q*

Synthese **99**: 137–172, 1994.

is irrational. If humans systematically fail to follow such a rule of logic, then humans are irrational. The experiments that are supposed to support the irrationality thesis (what I call *the irrationality experiments*) are said to provide evidence for such systematic failures.

My characterization of the rationality and irrationality theses is, so far, overly simple. No one thinks that humans *never* make mistakes in reasoning. Even the world's best logician might make errors if she has not had enough sleep. The rationality thesis claims that *under the right conditions* humans reason in accordance with the norms of rationality. The right conditions are those in which a person is able to reason at her optimal capacity: they are conditions under which a person reasons in accordance with her "cognitive competence."[3]

The notion of cognitive competence is based on the competence-performance distinction as used in linguistics. A person's linguistic competence is her underlying knowledge of language, her ability to understand and utter grammatical sentences.[4] People often make mistakes and, for example, utter *un*grammatical sentences. These errors are not, however, due to any deficiencies in a person's linguistic competence, but rather are due to nonlinguistic factors such as insufficient memory, lack of attention, high amounts of alcohol in the blood stream, and so forth. Failing to properly apply a rule of one's linguistic competence is called a *performance error*. The application of this distinction allows linguists to focus on the essential features of human linguistic capacity and to ignore the static of performance errors that often affect actual linguistic behavior.

The competence-performance distinction can be usefully applied to reasoning. People have an underlying ability to reason – cognitive competence – but sometimes, due to performance errors, they fail to reason in accordance with that ability. Defenders of the rationality thesis see all of the divergences from the norms of reasoning that humans make as performance errors. As such, they do not see these errors as indicative of an underlying ability to reason. Defenders of the *irrationality* thesis can agree that the competence-performance distinction is applicable to the realm of reasoning, but they deny that our cognitive competence matches the norms of reasoning. In short, the rationality thesis says that human cognitive competence matches the normative standards of reasoning (that is, the rules embodied in our cognitive competence are the same as those that we ought to follow), while the irrationality thesis denies this.

The argument for the rationality thesis which I focus on here says that human cognitive competence cannot diverge from the norms of reasoning because both the norms and competence are intimately connected with our intuitions about what constitutes good reasoning. As such, the argument requires interpreting the irrationality experiments – those psychological experiments that seem to provide evidence that humans are irrational – as not in fact providing evidence for the irrationality thesis. I shall be concerned with both the general version of this argument and a particular version of it – put forward by L. Jonathan Cohen – that draws on the theory of reflective equilibrium. Although this argument is innovative and *prima facie* plausible, it does not ultimately provide good reasons for believing that humans are rational and that the irrationality experiments are irrelevant to developing a characterization of human cognitive competence. In Section 2, I will sketch the general argument, examine why it might initially seem plausible, and argue that this initial plausibility is deceptive. In Section 3, I spell out the reflective equilibrium version of the argument for the rationality thesis. In Sections 4, 5, and 6, I turn to an extensive evaluation of this argument. In Section 7, I conclude.

2.

The argument for the rationality thesis that concerns me in this paper is a straightforward one. The basic idea of the argument is that human cognitive competence cannot diverge from the norms of reasoning because both are intimately connected with our intuitions about what constitutes good reasoning. In its simplest form, the argument is as follow:

(1) The normative standards of reasoning come from our intuitions about what constitutes good reasoning.

(2) Our intuitions about what constitutes good reasoning come from our cognitive competence.

(3) Therefore, the normative standards of reasoning come from our cognitive competence.

(4) Therefore, cognitive competence must match the normative standards of reasoning.

Versions of this argument have been defended by philosophers and psychologists alike. The argument, especially in this general form, seems sound: both premises (1) and (2) seem quite plausible; the argument from (1) and (2) to (3) is obviously valid because (3) follows from the two premises by the principle of transitivity; and the move from (3) to (4) seems reasonable.

Perhaps the primary reason why this argument for the rationality thesis seems so convincing is because of the apparent strength of the analogy between reasoning and language to which friends of the rationality thesis so frequently refer.[5] The comparison with language draws on the idea that we determine what linguistic competence is by studying our linguistic intuitions and also that we determine what the norms of linguistics are in the same way – there is no higher court of appeal as to what the rules of grammar are save our linguistic intuitions. The argument applied to language is as follows

(L1) The normative standards of grammaticality come from our intuitions about what constitutes grammaticality.

(L2) Our intuitions about what constitutes grammaticality come from our linguistic competence.

(L3) Therefore, the normative standards of grammaticality come from our linguistic competence.

(L4) Therefore, linguistic competence must match the normative standards of grammaticality.

(L1) and (L2) seem true and (L4) seems to follow directly from their truth. It is the strength of this argument that, by analogy, seems to lend support to the parallel argument for the rationality thesis. If the analogy between reasoning and language is strong, then the argument for the rationality thesis is in good shape.

Although there are many features that language and reasoning have in common, I do not think that the analogy can do the work required to make the argument for (4) a strong one. Suppose the human brain was constructed such that linguistic competence was altered in a small but nontrivial way: some basic linguistic patterns that are in fact judged grammatical, would, if some part of the brain that deals with language were constructed differently, be judged ungrammatical (call these type A patterns), and some linguistic patterns that are in fact judged ungrammatical would, if the brain were constructed differently, be judged grammatical (call these type B patterns). In such a case, both linguistic *competence* and linguistic *norms* would change. If our brain were constructed differently, not only would type A patterns be *judged* ungrammatical and type B patterns be *judged* grammatical, but also type A patterns would *be* ungrammatical and type B patterns would *be* grammatical. The point is that the grammaticality of linguistic patterns is dependent on the actual structure of the human brain.

Consider the same sort of case in reasoning. Suppose, as (4) claims, that human cognitive competence is appropriately characterized as matching the norms of reasoning. Now imagine that the brain were constructed differently and that certain heuristics which humans in fact follow were, as a result of the different brain structure, not followed (call these type A heuristics). Further, imagine that certain other heuristics which humans in fact do *not* follow were, as a result of the different brain structures, followed (call these type B heuristics). Given the assumption that humans are *in fact* rational, in the imagined counterfactual situation, humans are *not* rational. This is because, by embracing the rationality thesis, one is accepting that the principles of reasoning which are in fact embodied in our cognitive competence are the right principles. This further commits one to the claim that any *other* principles (for example, principles embodied in type B heuristics) are the *wrong* ones. This is different from language because linguistic norms (principles of grammaticality) are relative to actual linguistic competence, while norms of reasoning (principles of rationality) are *not* relative to actual cognitive competence. Even if in fact human cognitive competence matches the norms of reasoning, it does not do so because the norms are indexed to actual competence as they are in linguistics.[6] This difference between linguistic competence and its relation to grammaticality on the one hand and cognitive competence and its relation to rationality on the other is devastating to the analogy between linguistics and reasoning behind the argument for the rationality thesis, that is, (4).

This shows the weakness of the analogy between linguistics and reasoning that makes plausible the argument from (1) and (2) to (3) and (4). The example does not, so far as what I have said, blame the weakness of the argument for (4) on either the logical structure of the argument or one or both of the premises. The example does, however, undermine the plausibility of the argument. Although the normative standards match human competence in the realm of language, further argument is required to show that they do so in the realm of reasoning. Fortunately, the general argument for the rationality thesis need not rely on the analogy with linguistics; one can try to make premises (1) and (2) and the form of the argument plausible on other grounds. In the next section, I turn to just such an account.

3.

Cohen (1981) attempts to defend the rationality thesis with an argument of similar structure to the one discussed in Section 2 but that makes use of reflective equilibrium. The theory of reflective equilibrium is an account of justification that says a set of rules is justified in some domain if the rules provide a coherent and explicit characterization of our judgments about that domain. If some version of premises (1) and (2) can be defended independently of the failed analogy to linguistics, then the argument for the rationality thesis would be on a strong footing. The reflective equilibrium version of the argument for the rationality thesis attempts to meet this challenge as follows:

(RE1) The normative standards of reasoning come from a process of reflective equilibrium with our intuitions about what constitutes good reasoning as input.

(RE2) A descriptive theory of cognitive competence comes from a process of reflective equilibrium with our intuitions about what constitutes good reasoning as input.

(RE3) Therefore, because both come from the same process with the same inputs, cognitive competence must match the normative standards of reasoning.

This argument is similar in form to the previously discussed argument for the rationality thesis. Both arguments see the normative standards of reasoning and the proper account of cognitive competence as coming from intuitions about what counts as good reasoning and both conclude that the norms and cognitive competence must match. The reflective equilibrium argument for the rationality thesis is, in contrast, plausible independent of the analogy to the linguistics discussed in Section 2. In particular, the argument is plausible because (RE1) is based on a respectable epistemological theory, reflective equilibrium.

On the reflective equilibrium account of justification, to justify a set of principles that characterizes judgments in a given domain, one generates rules that conform to commonly accepted judgments. If one such rule sanctions judgments that do not conform to general practice, the rule is modified; if, however, such a modification would produce a rule that is intuitively unacceptable, then the judgment is rejected. This process may be circular, but, according to Nelson Goodman, it is a "virtuous" circle – rules and inferences are justified together by being brought into agreement. He writes:

> Principles of deductive inference are justified by their conformity with accepted deductive practice. Their validity depends upon accordance with the particular deductive inferences we actually make and sanction. If a rule yields inacceptable inferences, we drop it as invalid. Justification of general rules thus derives from judgments rejecting or accepting particular deductive inferences.
>
> This looks flagrantly circular. I have said that deductive inferences are justified by their conformity to valid general rules, and that general rules are justified by their conformity to valid inferences. But this circle is a virtuous one. The point is that rules and particular inferences alike are justified by being brought into agreement with each other. A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend. The process of justification is the delicate one of making mutual adjustments between rules and accepted inferences; and in the agreement achieved lies the only justification needed for either.
>
> All this applies equally well to induction. An inductive inference, too, is justified by conformity to general rules, and a general rule by conformity to accepted inductive practice. Predictions are justified if they conform to the valid canons of induction; and the canons are valid if they accurately codify accepted inductive practice. (1983, pp. 63–64)

Since the time of Goodman's formulation of reflective equilibrium, this method has been used to justify principles in other realms besides deduction and induction. Most notably, perhaps, is John Rawls's application of reflective equilibrium to moral theory.[7] According to Rawls, to develop a theory of ethics, we begin with a set of moral judgment – for example, judgments such as "It is wrong to torture babies". These judgments are collected with an eye toward producing a set of principles – for example, principles such as "Always do whatever will minimize the total amount of pain and suffering and maximize the total amount of happiness" – that not only underlie these judgments but also extend and systematize them. We begin by articulating our strongly held considered judgments and a set of principles that would fit with these convictions. Presumably, there will be discrepancies that arise – some judgments that follow from the principles will not be among our considered judgments and some of our considered judgments will not fit with the principles. We will then endeavor to eliminate these discrepancies by modifying some principles and retracting some judgments. Eventually, Rawls says, we will come on "principles which match our considered judgments duly pruned and adjusted" (1971, p. 20). He calls this *reflective* because "we know to what principles our judgments conform and the premises of their derivation" and *equilibrium* "because our principles and judgments coincide" (p. 20). The justification of moral principles cannot, according to Rawls, "be deduced from self-evident premises or conditions on principles; instead, its justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent view" (p. 21).

The strategy of the reflective equilibrium argument for the rationality thesis involves applying the reflective equilibrium account of how norms are justified to the realm of reasoning, (RE1). This seems a reasonable starting place for an account of how the norms of reasoning are justified. The next step of the argument is to defend a reflective equilibrium account of cognitive competence, (RE2). Combined with the reflective equilibrium account of norms, the reflective equilibrium account of competence seems to provide a strong defense of the rationality thesis. In Section 4, I will examine the reflective equilibrium account of norms; in Section 5, I turn to the reflective equilibrium account of competence; and, in Section 6, I examine the validity of the argument that moves from the two reflective equilibrium accounts to the rationality thesis.

4.

Recall the first premise in the reflective equilibrium argument for the rationality thesis:

(RE1) The normative standards of reasoning come from a process of reflective equilibrium with our intuitions about what constitutes good reasoning as input.

There is a general objection to this premise: the process of reflective equilibrium will count some principles of reasoning as normative that are not in fact the right principles. There is evidence to suggest that included among the inferential principles which will be in reflective equilibrium for people are principles we have good reason to think are *not* rational. This is exactly the sort of evidence provided by the irrationality experiments, experiments that seem to provide empirical support for the irrationality thesis. These experiments suggest that, even in the face of evidence and extensive briefing to the contrary, subjects continue to violate rules of logic and principles of probability.[8]

Another example of a principle that is likely to be in reflective equilibrium but that is not rational is the gambler's fallacy, a particular instance of which would be the belief that a long stretch of coin-flips that come up as heads increases the probability that the next coin-flip will be a tail.[9] Following this fallacy and failing to apply the principles of logic and probability properly, as do subjects in the irrationality experiments, are practices that would be in reflective equilibrium for many people. According to (RE1), if a principle is in reflective equilibrium, then it is justified, but it is absurd, for example, to think people are justified in reasoning in accordance with the gambler's fallacy. The general point is that the results of the irrationality experiments as well as the gambler's fallacy are supposed to count against the reflective equilibrium account of the justification of principles of reasoning because they are examples of heuristics that are *not* rational but *are* in reflective equilibrium. In the remainder of this section, I will examine the resources that friends of the reflective equilibrium account of norms have to avoid the objection that some of the principles that are in reflective equilibrium are not rational.

One possible reply open to friends of the reflective equilibrium model is to bite the bullet, that is, to say that what it means to be justified just *is* to be in reflective equilibrium.[10] For example, if the gambler's fallacy is in reflective equilibrium for someone, then it is justified for her. This seems a doomed strategy, for this fails to do justice to what justification is. If such a principle could be justified, then being justified seems a vacuous notion. An advocate of the bite-the-bullet strategy might try to defend this strategy by comparing reasoning to linguistics. In linguistics, such an advocate might point out, if a linguistic principle is the result of balancing judgments about what particular utterances are grammatical with judgments about what in general is grammatical (that is, if a principle is in reflective equilibrium), then the principle is part of linguistic competence. Even if this was the right picture of how linguistics works, it would not help support the bite-the-bullet strategy because, as I have pointed out in Section 2, linguistic norms are indexed to actual linguistic competence, but norms of reasoning are not indexed to actual cognitive competence. The analogy to linguistics will not help the bite-the-bullet strategy; without some such help, this strategy is a nonstarter.

Another possible response to the gambler's fallacy-type objection to the reflective equilibrium model of the justification of the norms of rationality is to argue that only *considered* intuitions are involved in the reflective equilibrium process. The implication of this move is that the sort of intuitions behind nonrational principles of reasoning are not *considered* intuitions but *naive* ones. The idea is to modify (RE1) in the following manner:

(RE1′) The normative standards of reasoning come from a process of reflective equilibrium with our *considered* intuitions about what constitutes good reasoning as input.

By preventing unconsidered intuitions from entering the reflective equilibrium process, the hope is that no nonrational principles of reasoning would end up being in reflective equilibrium. The idea of modifying the reflective equilibrium process to prevent nonrational principles from being in reflective equilibrium is promising; there are, however, several points to make about this suggested modification.

It is interesting to note that this suggested modification to (RE1) is not open to Cohen. By "intuition", he means "an immediate and untutored inclination, without evidence of inference" (1981, p. 318), to make a particular judgment. It is these immediate and untutored inclinations that Cohen sees as being the input to the reflective equilibrium process. His main reason for focusing on naive intuitions connects to what he thinks is the correct picture of linguistics and linguistic intuitions. Cohen's insistence that naive intuitions are the ones relevant to developing a theory of cognitive competence seems mistaken. As such, Cohen's hesitance to embrace (RE1′) as an attempt to save the reflective equilibrium account of norms from the objection at hand seems unwarranted.

First, the analogy with linguistics on which Cohen bases his focus on naive intuitions will not do the work he wants it to because linguists do in fact focus on *considered* intuitions. Consider the naive intuition that

> The girl whom the cat which the dog which the farmer owned chased scratched fled.

is ungrammatical. If you sit down and carefully consider the sentence, you will see that it *is* grammatical. Note that the core of the sentence is:

> The girl fled.

Which girl fled? The answer is: The girl whom the cat scratched. So, we now have:

> The girl whom the cat scratched fled.

But which cat scratched the girl? The answer is: The cat which the dog chased. We now have:

> The girl whom the cat which the dog chased scratched fled.

Finally, Which dog chased the cat? The answer is: The dog which the farmer owned. We can now see the grammaticality of the sentence

> The girl whom the cat which the dog which the farmer owned chased scratched fled.

Note, however, that it seems grammatical only on reflection. Even knowing that it is grammatical, each time I look at this sentence, it takes me a moment to reconvince myself of its grammaticality. Considered linguistic intuitions are relevant to developing an account of linguistic competence. The analogy Cohen tries to make to justify his emphasis on naive intuitions is that cognitive competence is like linguistic competence, but linguistic competence is accessible only through considered (linguistic) intuitions.

Because the picture of linguistics as only involving naive intuitions is mistaken, Cohen's insistence that the reflective equilibrium account of the norms of reasoning involves only naive intuitions is unsupported. Thus, the suggestion that only considered intuitions are involved in this reflective equilibrium process, (RE1′), seems motivated. Recall that the idea behind this suggestion was that focusing on considered intuitions would insure that only rational principles would be in reflective equilibrium. For example, while the gambler's fallacy might be in reflective equilibrium with naive intuitions as input to the balancing process, the motivation behind (RE1′) is that the gambler's fallacy would not be in reflective equilibrium with only considered intuitions as input.

The problem with this suggestion is that there is no particular reason to think that restricting the intuitions involved in reflective equilibrium just to considered intuitions will block the gambler's fallacy-type problems for the reflective equilibrium model. Many people who fall prey to the gambler's fallacy will presumably accept the fallacy even under careful consideration. The same is true for the results of the irrationality experiments: subjects in the experiments not only make systematic errors of reasoning, but they also sometimes stubbornly insist, even in the face of evidence to the contrary, that they are correct to reason as they do.

Part of the reason why narrowing the input to the reflective equilibrium process to just considered intuitions is unhelpful as a way to address the objection that nonrational principles will be in reflective equilibrium is that it is unclear what is involved in the process of considering intuitions. The suggestions that follow attempt to take what seems correct about modifying (RE1) to (RE1′) – namely, that the reflective equilibrium process of justifying norms of reasoning needs to be narrowed in response to, for example, the gambler's fallacy – while putting forward a more specific proposal for what sort of modification ought to take place.

There are three additional replies to this gambler's fallacy-type argument against the reflective equilibrium view. First, one might narrow the range of people whose intuitions count in the reflective equilibrium process. A set of inferential principles would be justified, on this view, if they were in equilibrium for some class of experts. Second, one might widen the scope of reflective equilibrium by considering a broader set of rules and judgments, namely, besides our inferential rules and judgments, we could include epistemological, metaphysical, and other types of rules and inferences. Third, the wide reflective equilibrium view could be combined with the expert view, resulting in the view that a set of inferential principles are justified if they are in wide reflective equilibrium for some class of experts.[11] I will consider each possibility in turn.

The first suggested modification to the reflective equilibrium view of the justification of principles of reasoning, known as the expert version of reflective equilibrium, is to say that a principle is established as a norm of reasoning if it is in reflective equilibrium for those people in a position to assess the relevant considerations. Drawing on Hilary Putnam's theory of the division of linguistic labor (1975), the expert reflective equilibrium view says that a principle is justified if it would be the result of reflective equilibrium performed by society's experts (Stich and Nisbett, 1980); Putnam's theory says that, in every speech community, there are terms

> whose associated "criteria" are known only to a subset of the speakers who acquire the

> terms, and whose use by the other speakers depends upon a structured cooperation between them and the speakers in the relevant subsets. (Putnam, 1975, p. 228)

This idea is to be applied in the realm of reasoning in the following fashion: some of the principles of rationality (which apply for all humans) are not the result of reflective equilibrium applied to everyone's judgments about what counts as good reasons, but rather are the result of reflective equilibrium applied by a certain subset of people – namely, the set of people who are the experts – to their own judgments. This would avoid the problem of the gambler's fallacy and similar principles being in reflective equilibrium and hence being deemed justified because the experts, for example, probability theorists, would not accept the gambler's fallacy in reflective equilibrium. The expert reflective equilibrium view seems to work well in deeming as justified the inferences we think (on reflection) are justified.

A response to this modification to the reflective equilibrium account is to say that the modification only works to avoid cases like the gambler's fallacy because the experts we consult are already known for their reliability when it comes to following principles we think are justified. But this, so the response goes, is a question-begging way to justify our inferential practice (Stich, 1990, p. 86). The challenge is to develop a general account of justification; appealing to those experts who follow principles that are justified without a general account of how to figure out who the experts are simply begs the question.

Stephen Stich and Richard Nisbett (1980, p. 201) suggest a general account of who the experts are; they say that the people who count as experts in the justification of some particular inference are those who the person making the inference thinks are the experts. So modified, the expert reflective equilibrium view is again open to the gambler's fallacy counterexample. Suppose, when it comes to gambling, I think the average Las Vegas compulsive gambler is the expert (after all, such people have a great deal of practice at gambling); such a person, however, may well believe the gambler's fallacy is justified. The gambler's fallacy would, in turn, be justified for me. It is not just that I would *believe* it is justified – I may or I may not – but that, on the Stich and Nisbett interpretation of the expert reflective equilibrium view, I *would* be justified. The problem is that often the people who are deferred to are no more justified in their beliefs than those who defer to them.[12] In a nutshell, the complaint against the expert modification to the reflective equilibrium account of how our principles of reasoning are justified is that there is no non-question-begging way to pick out the experts.

Friends of the reflective equilibrium account might plead guilty to this charge. It is circular to appeal to experts, but the reflective equilibrium account is unashamedly circular. Who counts as an expert is just another part of what is figured into the reflective equilibrium process. Just as some intuitions about which particular inferences count as good reasoning might be rejected in reflective equilibrium, some intuitions about which particular people count as experts might be rejected. Just as originally unintuitive principles of reasoning might be accepted in reflective equilibrium, some people who originally seemed nonexperts might be counted as experts in reflective equilibrium. This is circular, friends of reflective equilibrium would admit, but they would say that it is a virtuous circle.

Although friends of the expert version of reflective equilibrium have resources to respond to the charge of circularity, it is not clear that they have adequate resources to respond to the original objection to the reflective equilibrium account of the norms of reasoning. This original objection is that reflective equilibrium will count some principles that are not in fact rational as norms of reasoning. This objection can rear its head again with respect to the expert modification in one of two ways, depending on who counts as an expert. If an expert is someone for whom the principles that are in reflective equilibrium are always rational, then the objection to the expert view is that reflective equilibrium may deem as an expert someone who is not in fact an expert. If an expert is *not* someone who is always correct about the norms of reasoning, that is, if an expert might accept a nonrational principle in reflective equilibrium, then the objection to the expert view is that those people who are deemed experts might deem rational principles that are in fact *not* rational. The choice between these two different accounts is a forced option – either an expert is always correct in reflective equilibrium or not – and either way, the expert view is open to the same objection that counted against the unmodified reflective equilibrium account.

The second suggested modification to the reflective equilibrium view of the justification of principles of reasoning is to say that a principle is established as a norm of reasoning if it is in *wide* reflective equilibrium.[13] Rawls, who coined the term "reflective equilibrium", makes a distinction between wide and narrow reflective equilibrium (1974/1975).[14] *Narrow* reflective equilibrium is achieved when a set of judgments is coherently systematized by (that is, brought into balance with) a set of general principles. This would be accomplished in ethics, for example, if we produced a set of principles from which all and only our somewhat altered and refined first-order moral judgments followed. The result of narrow reflective equilibrium is a coherent systematization of our moral judgments. *Wide* reflective equilibrium is achieved when a set of judgments, a set of principles, *and* a set of general philosophical theories (theories of personal identity, metaphysics, the social role of moral and political theory, etc.) are brought into agreement. The search for wide reflective equilibrium begins as does the search for narrow reflective equilibrium, but once our judgments and a set of general principles are brought into agreement, various alternative sets of balanced judgments and general principles are considered and these alternatives are then brought into balance with philosophical theories through the same sort of balancing process. This process, rather than producing a systematization of our judgments, is more revisionary. The set of principles that results from wide reflective equilibrium has a broader network of support and a reflective philosophical backing. As a result, there is a greater likelihood that wide reflective equilibrium will produce a theory that diverges from intuitions.[15]

As an example of wide reflective equilibrium, consider Derek Parfit's argument for utilitarianism (1984). Utilitarianism is the view that one ought to do whatever will cause the greatest amount of happiness and the least amount of unhappiness. A standard objection to utilitarianism is that it is not acceptable to balance losses and gains between people; to the extent that we have intuitions in favor of utilitarianism, these intuitions should be outweighed by our strong intuitions against interpersonal balancing. One set of intuitions that counts against interpersonal balancing is the separateness of persons. The separateness of persons is the claim that people are separate beings, each with his or her own life to lead; as such, people are the relevant units for moral theory. This is an objection to utilitarian theory because utilitarianism sees an important role for *inter*personal balancing. Intuitions about the separateness of persons suggests that utilitarianism will not be in narrow reflective equilibrium.

Parfit's response to these objections can be seen as fitting the wide reflective equilibrium model. Parfit can grant that the separateness of persons is an objection to utilitarianism in *narrow* reflective equilibrium. Parfit, however, produces arguments from metaphysics to the effect that persons are not the relevant units for moral theory. Identity of persons, Parfit argues, is not "what matters" to moral theory. Rather, according to Parfit, psychological continuity and connectedness are. But because I may be psychologically connected to other people besides myself, benefits and harms, pleasures and pains can, contrary to the separateness-of-persons objection to utilitarianism, be balanced among various people. If benefits and harms can be balanced in this fashion, our original intuitions against utilitarianism should be revised in the face of Parfit's arguments in metaphysics to the effect that persons are not the relevant units for moral theory. Parfit can be seen as arguing that utilitarianism is the result of wide reflective equilibrium applied to ethical theory.

I do not mean to endorse Parfit's conclusion in favor of utilitarianism; I just cite it as an example of wide reflective equilibrium. In fact, following Rawls's (1974/1975, Sect. IV) discussion of the relationship of moral theory to philosophy of mind and metaphysics, I am not clear whether the metaphysical conclusions that Parfit defends ought to count against the narrow reflective equilibrium conclusions in moral theory (namely, that utilitarianism is wrong) or whether the strength of our intuitions against utilitarianism (stemming, for example, from the strength of our intuitions in favor of the separateness of persons) ought to count against Parfit's metaphysical view in *wide* reflective equilibrium. This is obviously not the place to settle this issue; the relevant point is that settling it would be part of the wide reflective equilibrium process of bringing our moral intuitions, the moral principles that are relevant to these intuitions, and various philosophical arguments into agreement.

Returning to reasoning, the idea of appealing to the notion of wide reflective equilibrium is to argue that the gambler's fallacy and the principles suggested by the results of the irrationality experiments would not be in *wide* reflective equilibrium even if they are in narrow reflective equilibrium; such principles, says the defender of the wide reflective equilibrium view, would be rejected as a result of the process of balancing general principles of reasoning with philosophical and other theoretical considerations. For example, if people were persuaded by theoretical arguments in favor of standard probability theory, they would see that the gambler's fallacy is in fact a fallacy. This line of thought can be seen as suggesting the following modification to (RE1):

> (RE1″) The normative standards of reasoning come from a process of *wide* reflective equilibrium with our intuitions about what constitutes good reasoning as input.

Just as the other two modifications to the reflective equilibrium account of norm – the expert view and the considered intuitions view – the wide reflective equilibrium account attempts to prevent nonrational principles of reasoning from being deemed rational by the reflective equilibrium account.

A virtue of the wide reflective equilibrium account is that it makes clear why the norms of rationality are not indexed to cognitive competence. Recall from Section 2 that there is an important disanalogy between linguistics and reasoning – linguistic norms are indexed to linguistic competence while norms of reasoning are *not* indexed to cognitive competence. If the wide reflective equilibrium account of the norms of reasoning is right, this disanalogy is explained. According to the wide reflective equilibrium account, the norms of reasoning are the result of bringing into balance our inferential practices, our intuitions about what counts as good reasoning, and – this is the crucial part of the picture – general philosophical and theoretical considerations. Because theoretical considerations are brought in, a wide reflective equilibrium account can be highly revisionary with respect to our original intuitions and practices; the result is that our intuitions and naive practices can be dispensed with in wide reflective equilibrium. Norms of reasoning are thus not indexed to cognitive competence. This is in contrast to linguistics. In linguistics, general philosophical considerations are not brought into the process of determining the linguistic norms; the process of developing linguistic norms is *not* highly revisionary and thus *is* indexed to linguistic competence. (RE1″) thus has the virtue of fitting with an important fact about the justification of norms of reasoning.

Against (RE1″), Stich (1990, p. 85) has pointed out that it is difficult to assess whether, for example, a person who actually accepts and follows the gambler's fallacy will give up this principle in the face of philosophical considerations against it (after all, a gambler is likely to give much less weight to some 'bookish' philosophical principle than to a principle she has 'learned to trust' after years of experience in casinos). He goes on to argue that it is possible for a person to settle on a wide reflective equilibrium that includes 'some quite daffy' (p. 86) rules of inference.

One possible reply to this argument (the third modification to the reflective equilibrium view) is the expert wide reflective equilibrium view which combines the first two modifications. On this view, a principle is justified if it is the result of experts engaging in the process of wide reflective equilibrium. This view might be seen as an improvement to the wide reflective equilibrium view for it might seem to reduce the likelihood that an unjustified principle will be in wide reflective equilibrium. The same problem, however, remains for the expert wide view; even if the chances are reduced that an unjustified principle will be deemed justified, it remains *possible* for this to happen.

There are some interesting and potentially strong defenses of the wide reflective equilibrium account (in both its expert and its nonexpert versions) against Stich's objection that, even in wide reflective equilibrium, one (even an expert) might embrace a "daffy" inferential principle. The problem is that Stich is not clear about what he means by "daffy" in this argument. If by a "daffy" principle, he means a principle that we would currently judge not to be justified, then surely it is true that a daffy principle might turn out to be justified on the wide reflective equilibrium view. But this is not an objection to wide reflective equilibrium. Widening the scope of reflective equilibrium allows for the possibility that certain principles that naively seem unjustified (that is, daffy) will, when various philosophical considerations are presented, be seen to be justified after all (that is, *not* daffy). While it is a result of the wide reflective equilibrium picture that principles we currently think of as daffy will be justified, this fact is not an objection to the view. No doubt, some of the principles we currently think of as daffy are justified; whatever method actually justifies principles of reasoning, *some* of the principles of reasoning that we currently reject ought to be accepted on a good account of justification.

On the other hand, by a "daffy" principle, Stich might mean an *objectively* daffy principle, that is, a principle which is not in fact a normative principle of reasoning (whether or not we *think* it is a norm). On this reading of the term, a principle that is (objectively) daffy could be the result of wide reflective equilibrium. Simply put, this objection against wide reflective equilibrium is that the principles that wide reflective equilibrium deems are rational might not, in fact, be rational;

as such, wide reflective equilibrium is not an adequate account of how normative principles are justified.

I will briefly sketch two possible replies to this objection. The first – and perhaps the most radical – reply is to embrace a coherence theory of truth, that is, one which says that the true complete theory of the world is the most coherent collection of beliefs about the world. On this account of what it means to be true, a particular belief is true if and only if it fits with the maximally coherent theory of the world. Similarly, on this theory, a principle would be rational if and only if it is among the maximally coherent set of principles. Because the process of wide reflective equilibrium arguably produces a maximally coherent set of principles, the principles justified by wide reflective equilibrium would (contra Stich) be guaranteed to be the most rational. This is not the place to spell out all of the problems and disadvantages of this view, but I will just mention that perhaps its most serious consequence is that it is incompatible (or, at least, is in tension) with the most *prima facie* plausible metaphysical theory, namely, realism. This consequence alone may involve too high a price to pay for defending a wide reflective equilibrium account of justification.

Another reply to the argument that wide reflective equilibrium will deem justified some principles which in fact are *not* rational would be to grant that it is possible for the principles justified by wide reflective equilibrium to diverge from the rational principles while denying that this is a serious criticism of wide reflective equilibrium. This strategy might be motivated by pragmatic considerations; namely, by denying that there is any strategy for justifying principles of reasoning that could do better than wide reflective equilibrium. On this view, wide reflective equilibrium tells us what people in the human epistemological position are justified in believing and what principles people in the human epistemological position are justified in following. Perhaps what humans are justified in believing is not in fact true and perhaps the principles humans are justified in following are not in fact justified, but there is no particular reason to think that they are not and no better way of figuring what beliefs are true and what principles are justified other than wide reflective equilibrium.

This is not the place to mount a complete defense of a reflective equilibrium account of the norms of reasoning. Suffice to say that, insofar as reflective equilibrium is an interesting epistemological theory, it also seems a good theory of the justification of the norms of reasoning, particularly because there do not seem to be any serious competitors. Friends of the reflective equilibrium account have ample resources to employ in answering the primary objection to their favored account. Narrowing the range of people whose balancing of judgments is relevant to justification (the expert modification) and expanding the inputs to the balancing process to include broader theoretical considerations (the wide modification, (RE1″) seem promising strategies to prevent non-rational principles from being counted as rational ones. For my purpose of assessing the reflective equilibrium argument for the rationality thesis, some version of (RE1) is plausible enough both to make the reflective equilibrium argument interesting and to suggest that I should turn my attention to other parts of the argument.[16] In the next section, I evaluate (RE2).

5.

The second premise in the reflective equilibrium argument for the rationality thesis is an account of how human cognitive competence should be researched:

(RE2) A descriptive theory of cognitive competence comes from a process of reflective equilibrium with our intuitions about what constitutes good reasoning as input.

On this account, psychologists start their investigation of cognitive competence by looking at the reasoning behavior of particular individuals. From this behavior, researchers attempt to develop a generalized characterization of human reasoning ability, namely, a set of rules that approximately fit actual inferential behavior. These rules would not perfectly characterize the observed behavior because some of this behavior would be the result of interference with the operation of cognitive competence, that is, performance errors. Researchers then ask individuals whether they think they are following these rules of inference. If the individual identifies a rule as one she uses and given that it accords with her behavior, the rule is accepted; otherwise, if the rule is not accepted by an individual reasoner, it is rejected as a norm unless it is strongly supported by her behavior.

To see the plausibility of this, imagine trying to learn the rules of chess just by watching people play chess. You would just watch the moves that people make and then try to abstract the rules of the game from these moves. These rules would then be used to make predictions about what people will do in various situations, predictions that you can test by further observing chess games. Because you do not know the rules of chess when you start, you will not initially be able to determine if someone has made an illegal move; you will simply note the moves without being able to distinguish between the legal and illegal ones. When you try to abstract from the behavior of players, you may find it difficult to generate any rules that fit with all of the observations, except those that are very complex. For example, suppose you note that a certain shaped chess piece (what those of us in the know call a rook) always moves horizontally or vertically, except on one occasion when you observed someone move a rook diagonally. The rule 'Rooks can move vertically, horizontally, and diagonally' would fit with all the observations you have made, but it might seem odd, if this is one of the actual rules, that only one person in all the games you have watched has taken advantage of the rook's ability to move diagonally. Further, the part of the rule that sanctions diagonal moves would not help you to make any additional correct predictions (unless, of course, someone made the same sort of illegal move again). The rule 'Rooks can move vertically and horizontally, unless it is 3 September 1990 [the day you observed the rook being moved diagonally], in which case it can also move diagonally' would also fit with all the data. None of the other rules of the game seems however, to be indexed to a particular date. Instead, a sensible strategy would be to throw out the aberrant rook move as some sort of a performance error and opt for the rule 'Rooks can move vertically or horizontally'. This would be a rule of competence.[17]

According to the reflective equilibrium account, the study of human cognitive competence involves a similar process. The rules that characterize cognitive competence are not directly accessible to cognitive scientists in the same ways that the rules of chess are not directly accessible to a naive observer of chess. In both cases, the observers must start by looking at behavior (chess-playing behavior on the one hand and reasoning behavior on the other). In the chess example, however, it may be more difficult to generate idealized, abstract rules through observation alone than it is to discover a person's cognitive competence. The strategy of observe, abstract, idealize can be supplemented in the case of human reasoning by asking the subject whether a specifically chosen inference is valid or whether a particular rule fits with her intuitions. From just observing behavior, we might develop a rule of performance that says to infer *q* from *p implies q* and *p* except when very drunk or very tired. A person would not agree to this rule because she believes that *p* and *p implies q* together entail *q* regardless of the amount of alcohol or sleep any person has had; instead, she would accept the idealized version of the rule (that is, infer *q* from *p implies q* and *p*) as a rule of competence. The strategy for developing rules that characterize cognitive competence is thus observe, abstract, idealize, test, revise, test, revise, and so forth. This is roughly the picture suggested in (RE2).

In the remainder of this section, I shall argue that the reflective equilibrium account of cognitive competence is false. (RE2) offers this as an account of how cognitive psychology ought to be done, in particular, as a theory of how people's underlying ability to reason ought to be characterized. There are two general reasons why this account is wrong: the account gives too much weight either to naive intuitions or to introspection and it ignores other sorts of data relevant to developing a theory of cognitive competence.

My first objection to (RE2) is that people's intuitions about their own cognitive competence should not be given the central role the reflective equilibrium account gives them. By "intuition", Cohen means "an immediate and untutored inclination, without evidence of inference" (1981, p. 318), to make a particular judgment. He seems to think that intuitions, so defined, are central to a theory of cognitive competence. In Section 4, I argued that this emphasis on naive rather than considered intuitions is mistaken because it is based on an analogy with an erroneous picture of linguistics; Cohen mistakenly assumes that linguistic competence is based on naive intuitions, but it is really based on considered ones.

Cohen's insistence that naive intuitions are the ones relevant to developing a theory of cognitive competence is problematic for another reason. By focusing on naive intuitions, just the sorts of mistakes friends of the reflective equilibrium argument for the rationality thesis want to deem performance errors, that is the mistakes pointed out by the irrationality experiments, will inevitably be included in cognitive competence. For example, if psychologists observe people's betting behavior, they will see that people bet in accordance with the gambler's fallacy. People's naive intuitions will tend to agree with this principle of reasoning. The gambler's fallacy is thus likely to be included in an account of cognitive competence that results from bringing into reflective equilibrium a person's naive intuitions about what principles she is following. This is because people admit they are following the gambler's fallacy (but they often do not admit that, in following this principle of gambling, they are reasoning fallaciously).

These two considerations suggest that, to be successful, the reflective equilibrium argument ought to focus on considered intuitions.[18] The suggestion is that our considered intuitions about what counts as good reasoning are to be taken into consideration in the development of an account of actual human cognitive competence, namely:

(RE2′) A descriptive theory of cognitive competence comes from a process of reflective equilibrium with our *considered* intuitions about what constitutes good reasoning as input.

This modification moves the reflective equilibrium argument toward considering reflection on one's own cognitive processes as a source of data for cognitive competence. Cognitive scientists should not observe behavior and then compare the principles they abstract from this behavior to people's naive intuitions about what principles they are following. Instead, they should compare the principles that characterize observed reasoning behavior with people's carefully considered intuitions as to what principles they are following. This sort of careful self-examination of what principles one is following is called *introspection*. Introspection is a method of research with a long tradition in psychology. It is a method, however, from which Cohen (1981, p. 318) explicitly distances himself. Perhaps Cohen's refusal to embrace introspection is justified because the move to introspection is problematic.

In 1879, when Wilhelm Wundt set up the first psychology laboratory, introspection was *the* method of research. Subjects in Wundt's laboratory were trained to report their own cognitive processes under experimental conditions. One reason why introspection is not the research method of choice in cognitive science is that dozens of psychological experiments have shown that introspective reports about human psychology are wrong.[19] One classic experiment of this sort, reported by Saul Sternberg (1966), involves giving subjects a list of randomly chosen single-digit numbers to memorize and then timing them to see how long they take to indicate whether a particular number is on the memorized list. For example, a subject might be shown the list '4 2 7 9 6' and asked whether the number nine is on the list. Sternberg found that subjects' reaction times (the amount of time it takes a subject to determine whether a number is on the list) vary with the length of the list but do *not* vary with the number's position on the list. This suggests that subjects are searching the list number by number and that the searching through the list continues even after the number has been found earlier in the search. For example, if the subject has been shown the list '4 2 7 9 6' and is asked whether the number two is on the list, she might (subconsciously) first look at the four and ask 'Is this a two?' and so forth. Further, she might continue searching through the list until the end, examining seven, nine, and six and asking if each is a two, even though the number two has already been found to be on the list.[20] That this is the heuristic we use to determine if a number is on a list is highly counterintuitive. It is much more intuitive that we stop looking through the list once we find the number we are looking for. It seems highly unlikely that an observe, abstract, idealize, etc., process using our considered intuitions as data would produce this description of our cognitive competence; even careful introspection would be unable to discover that this heuristic governs our list-searching behavior. The reflective equilibrium account of cognitive competence, as so far characterized, does not fit with the facts of what we actually know about human cognitive processes.

Further reason that our cognitive heuristics may be inaccessible to introspection (not to mention to untutored intuitions) can be shown by examining linguistics. Although humans can utter and comprehend highly structured, complex linguistic sentences, most nonlinguists have little understanding of how we perform the linguistic feats we do. Most of us have almost no intuitive sense of underlying linguistic structure and how it works. For example, consider the sentence 'John saw Bill's father shoot himself'. All native English speakers are able to interpret this sentence unambiguously as meaning that Bill's father was the one who got shot, not John or Bill, but complex linguistic theories need to be brought in to explain why. We all have *implicit* understanding of abstract linguistic principles, but few people (if any) have any *conscious* understanding of these underlying linguistic principles; further, those who do understand them get their knowledge from years of study, not simply from introspection. The upshot of this most recent discussion is that a reflective equilibrium account of cognitive competence that takes only considered intuitions as input, (RE2′), does not do justice to the way cognitive scientists should or actually do study human cognitive competence.

The second, and more serious, problem with the reflective equilibrium account of cognitive competence is that there are several *other* sources of data besides behavior, intuitions (naive or considered), and introspection that psychologists can and do make use of including, for example, neurophysiology, theory of computation, and evolutionary theory. To see this, suppose neuroscientific research advances in such a fashion that neuroscientists can isolate cognitive mechanisms in the brain. Although such research may be quite far from the current state of neuroscience, it is possible to make such discoveries. If they were made, these results would surely be relevant to a theory of cognitive competence. That, for example, some particular set of neurons is responsible for the application of *modus ponens* would count as support for the view that cognitive competence includes the ability to apply *modus ponens*. Note that such data is not derived from behavior or from introspection. In the realm of the technologically more realistic, Christopher Cherniak (1988) has argued that basic neuroscientific facts such as the size of the brain and the speed at which neurons operate are quite important considerations for the development of a theory of cognitive competence. He points out that, given the size of the brain, the number of neurons in it, the speed at which neurons operate, and the time it takes a human to make a calculation, there are many seemingly plausible cognitive heuristics that are *not* realizable in the human brain. Neuroscientific data is thus relevant to cognitive science and the development of a theory of cognitive competence; however, it is *not* the sort of data that is deemed relevant by the (RE2) and (RE2′) accounts of how cognitive competence is researched.

In addition to neuroscience, evolutionary theory is relevant to the study of cognitive competence. Leda Cosmides (1989) offers an account of deductive reasoning that is influenced by the constraints that evolution and natural selection place on psychological mechanisms. She writes:

> Natural selection, in a particular ecological setting, constrains which kinds of traits can evolve. For many domains of human activity, evolutionary biology can be used to determine what kind of psychological mechanisms would have been quickly selected out, and what kind were likely to have become universal and species-typical. Natural selection therefore constitutes "valid constraints on the way the world is structured"; hence,

> knowledge of natural selection can be used to create computational theories of adaptive information-processing problems. Natural selection theory allows one to pinpoint adaptive problems that the human mind must be able to solve with special efficiency, and it suggests design features that any mechanism capable of solving these problems must have. (1989, p. 189).

The point is that because certain adaptive problems are likely to be important for survival and, hence, to be selectively advantageous, certain cognitive heuristics will be more likely to have evolved given the evolutionary history of human beings. Evolutionary considerations can thereby inform research into cognitive competence by suggesting which heuristics are evolutionarily feasible and probable. The reflective equilibrium account of the study of cognitive competence has no room for such considerations.

Finally, Cherniak (1986) argues that computational theory is relevant to developing a theory of cognitive competence. Heuristics that might seem plausible candidates for being part of human cognitive competence cannot be implemented in the amount of time that humans in fact take to make certain inferences. For example, Cherniak (1986, pp. 47–54) criticizes Quine's "web of belief" model (Quine, 1961; Quine and Ullian, 1970) as requiring particular heuristics – that is, heuristics for belief modification, heuristics for testing consistency, and so forth – that are far too computationally demanding for the human brain to handle in a reasonable amount of time.[21] Computational considerations thus provide constraints on developing an account of cognitive competence, but such considerations are not available to the reflective equilibrium account.

That these three examples of types of data are relevant to the study of cognitive competence counts against (RE2) or (RE2′). This conclusion should be of no surprise because (RE2) gets its initial plausibility from an analogy with linguistics, a research program that includes study of more types of data than just actual linguistic behavior and linguistic intuition; neuroscience (Caplan and Hildebrandt, 1988), theory of computation (Wexler and Culicover, 1980; Berwick and Weinberg ,1980), and, perhaps, evolutionary theory (Pinker and Bloom, 1990)[22] are relevant to linguistics as well. This is not to say that linguistic behavior and linguistic intuitions are not relevant to the study of linguistic competence. Nor is it to say that inferential behavior and intuitions about what counts as good reasoning are not relevant to the study of cognitive competence. It is, however, to say that these considerations are not the only evidence relevant to the study of cognitive competence.

These two objections to the reflective equilibrium account of cognitive competence – that the account cannot explain actual advances in the understanding of cognitive competence and that other considerations besides behavior and intuitions are relevant to cognitive competence – are connected. The reason why the reflective equilibrium account of cognitive competence is impoverished is because it does not include the variety of other considerations that are relevant to cognitive competence. The reflective equilibrium account of cognitive competence is mistaken; (RE2) and (RE2′) are both false.

6.

Recall the reflective equilibrium argument for the rationality thesis:

(RE1) The normative standards of reasoning come from a process of reflective equilibrium with our intuitions about what constitutes good reasoning as input.

(RE2) A descriptive theory of cognitive competence comes from a process of reflective equilibrium with our intuitions about what constitutes good reasoning as input.

(RE3) Therefore, because both come from the same process with the same inputs, cognitive competence must match the normative standards of reasoning.

In Section 4, I defended the plausibility of (RE1) appropriately modified; some version of the wide reflective equilibrium account of the norms of reasoning is (at least) plausible. If the balancing process is broadened to consider general theoretical considerations, if the class of people with relevant balancing processes is narrowed, and so forth, then it seems the objection that the reflective equilibrium process justifies some irrational principles may be answered or at least the objection may be shown to be irrelevant. In Section 5, I argued that (RE2) was not at all so plausible. The reflective equilibrium account of cognitive competence is mistaken. Both our *considered* intuitions and evidence from various scientific disciplines are relevant to cognitive competence.

A friend of the reflective equilibrium argument for the rationality thesis might attempt to modify (RE2). The aim of such a modification would be twofold. First, this modification would be designed to defend a reflective equilibrium account of cognitive competence that is true, in particular, that takes the objections of Section 5 into account. Second, this modification would be designed to defend a reflective equilibrium account of cognitive competence according to which the inputs to the reflective equilibrium process and the reflective equilibrium process itself were the same with respect to the norms of reasoning. In this section, I will sketch a modification to the reflective equilibrium account of cognitive competence, (RE2), that avoids the objections of Section 5 and that matches the reflective equilibrium account of the norms of reasoning. I will argue that even this attempt to save the reflective equilibrium argument for the rationality thesis fails.

My candidate modification is to see the study of cognitive competence as a process of bringing our considered intuitions about reasoning into reflective equilibrium with our advanced scientific theories. On this picture, our considered intuitions about reasoning would be brought into balance with the relevant scientific theories – for example, neuroscientific, psychological, evolutionary, and computational theories. This picture of how a descriptive theory of cognitive competence is studied is similar to the picture I painted at the end of Section 5 of how linguistic competence is properly studied. Linguists do make use of (considered) linguistic intuitions, but scientific data (for example, neurophysiology, computational theory, and perhaps evolutionary theory) are relevant as well. Further, the relationship between our linguistic intuitions and the scientific data relevant to linguistics does seem to fit the sort of balancing involved in reflective equilibrium. While this picture diverges from Cohen's picture of cognitive competence, it does fit with the reflective equilibrium account. Further, it explicitly answers the objections I raised in Section 5. The suggestion is that (RE2) be modified as follows:

(RE2″) A descriptive theory of cognitive competence comes from a process of reflective equilibrium with both our intuitions about what constitutes good reasoning and scientific knowledge relevant to what constitutes good reasoning as input.

The crucial question for the argument for the rationality thesis is whether this account of cognitive competence is parallel to the reflective equilibrium account of the norms in such a way that the reflective equilibrium argument for the rationality thesis will go through.

For the reflective equilibrium argument for the rationality thesis to work, the reflective equilibrium process involved in cognitive competence must have the same data as input. For the reflective equilibrium account of cognitive competence to be true, it must include scientific evidence as input. It seems, however, that, for the reflective equilibrium account of norms to be true, it must not include scientific evidence as input. Scientific evidence – in particular, physical, chemical, psychological, and other facts about the brain – should play a role in the development of a descriptive theory of cognitive competence. But how can such facts play a role in the development of a normative theory of reasoning? Any such attempt seems to be guilty of the naturalistic fallacy of deriving 'ought' from 'is'.

Many, however, have suggested that the naturalistic fallacy is not a fallacy at all, that philosophical questions (most notably, epistemological and ethical questions) can and ought to be 'naturalized'. Perhaps rationality should be naturalized as well. If this is correct, then it may be perfectly acceptable for the reflective equilibrium process that determines the norms of reasoning to include scientific evidence as input. In fact, included as part of the very scientific evidence relevant to 'naturalizing rationality' would be the results of the irrationality experiments. This is all well and good as far as some of us may be concerned, but it will not be of help to friends of the rationality thesis, for they want to insulate human rationality from the potentially damning empirical evidence of the irrationality experiments. Advocates of the reflective equilibrium argument for the rationality thesis want to discount the evidence resulting from experiments such as the selection task and the conjunction experiment by saying it merely indicates the sorts of performance errors humans typically make – it does not illuminate our cognitive competence. So this sort of evidence is *not* available to them as part of the project of naturalizing rationality. Such evidence is, however, just the sort of evidence that bears on a descriptive theory of cognitive competence once we realize that empirical evidence *is* relevant to such a descriptive theory. This shows that the inputs to the reflective equilibrium processes of developing both a descriptive and a normative account of cognitive competence are different; this, in turn, blocks a reflective equilibrium defense of the rationality thesis because such a defense turns on there being the same inputs to both reflective equilibrium processes. Even if (RE2″) is true, the reflective equilibrium process involved in determining the norms includes inputs that are different from those the reflective equilibrium process for determining cognitive competence does.

There is, however, a further problem for the reflective equilibrium argument for the rationality thesis: it is unclear that the process involved in developing a theory of cognitive competence and the one involved in determining the norms of rationality would be the same *even if* their inputs were the same. The goal of the first process would be to develop a psychological account of human competence in reasoning and the goal of the second is to develop a normative account of reasoning. Even if the reflective equilibrium model of developing a descriptive theory of cognitive competence, (RE2″), and some version of the reflective equilibrium account of justification are true, and even if the reflective equilibrium processes get the same data as input, there is no reason to think the balancing process involved in developing a psychological theory would parallel the balancing involved in justification. In particular, in light of the different goals, the inputs to the two processes, even if they are the same, would be weighted in different ways as part of the balancing process. Even if an intuition is part of the input to both the reflective equilibrium process for determining the norms of reasoning and the reflective equilibrium process for determining human cognitive competence, this intuition will carry different weight in the two different processes. The same is true with a scientific fact such as that the brain contains a certain number of neurons. Even if such a fact is part of the input to both the reflective equilibrium process for determining the norms of reasoning (assuming the legitimacy of naturalizing rationality) and the reflective equilibrium process for determining human cognitive competence, there is good reason to think that this fact would be relevant to the outcome of the two reflective equilibrium processes in different ways. Given that the inputs (even if they are the same) will probably be weighted in different ways because of the different goals of the two reflective equilibrium processes, the outcome of the two processes will probably diverge. This counts against the reflective equilibrium argument for the rationality thesis.

This point may be made clearer by considering an example from a different realm. Consider once again the application of reflective equilibrium to ethics. Suppose that the reflective equilibrium model is applied to both the project of determining what is moral and the project of determining what moral sentiments human have. Further, suppose that the inputs to the two processes are the same. Given all this, a particular input to the reflective equilibrium process, say, for example, the intuition that it is wrong to torture babies, will be weighted in a particular way and will interact in a particular way with other inputs as part of the project of determining what is moral that will almost surely differ from the way the same intuition is weighted and interacts as part of determining what human moral sentiments are. As such, the results of the two reflective equilibrium processes will be different.

The attempt to modify the reflective equilibrium account of cognitive competence fails because it must include scientific evidence in the input to the balancing process. If one makes this modification and includes scientific evidence, the input to the process involved with cognitive competence and the process involved with the norms of reasoning will differ. If both processes do not have the same input, the reflective equilibrium argument for the rationality thesis fails. Even with the same input, however, the argument fails because the input will be weighted in different ways given the different goals of the two reflective equilibrium processes.

7.

The reflective equilibrium argument for the rationality thesis turns on there being an isomorphism between how norms of rationality are justified and how a theory of cognitive competence is developed. This isomorphism fails to hold. Cohen's version of the reflective equilibrium argument tries to establish this isomorphism by arguing that both processes fit the narrow reflective equilibrium model of justification. I have argued that neither process is appropriately characterized by narrow reflective equilibrium. With respect to the study of cognitive competence, the reflective equilibrium process involves scientific evidence as input. With respect to the norms of reasoning, the process involved is wide reflective equilibrium. Further, I have argued that an attempt to defend the required isomorphism based on both processes fitting a wide reflective equilibrium model fails as well because even if the two processes do fit the same model (which is far from obvious), they have different inputs, and even if they did have the same inputs, they have different goals.

The conclusion of this essay is, in a sense, no surprise, given the point I made in Section 2 against the general argument for the rationality thesis. My point, against the analogy between developing a theory of linguistic competence and developing a theory of cognitive competence, was that norms of grammaticality are indexed to actual facts about human psychology, neurophysiology, and the like, whereas norms of reasoning are not. Because the process of developing a descriptive theory of cognitive competence is, regardless of whether or not it involves either wide or narrow reflective equilibrium, indexed to empirical facts about humans, we should expect the results of such a process to diverge from a normative theory of cognitive competence.

Norman Daniels (1980a) has made an interesting and somewhat parallel point. He argues that the analogy suggested by Rawls (1971, pp. 46–48) between ethics and linguistics is mistaken. Linguistics, Daniels argues, involves *narrow* reflective equilibrium while ethics involves *wide* reflective equilibrium. While I am not completely convinced that linguistics and cognitive science are appropriately characterized as *narrow* reflective equilibrium, I think that Daniels is on the right track in pointing to the difference between ethics and linguistics. Ethical principles – *like* principles of reasoning and *unlike* linguistic principles and psychological descriptions – are independent of physiological facts about humans. Justifying principles of ethics or reasoning involves general philosophical reflection in a way that justifying psychological or linguistic principles does not. So if Daniels is correct about ethics and I am correct in thinking that justifying principles of reasoning is like justifying principles of ethics in the relevant ways, even if a wide reflective equilibrium account of the justification of principles of reasoning can be developed, the analogy between this process and the psychological project of determining actual human cognitive competence does not hold.

Despite the failure of the reflective equilibrium argument to establish that human cognitive competence matches the norms of reasoning, human cognitive competence *might* still match the norms of reasoning.[23] In light of the results of the irrationality experiments and failing a good argument that these results should be viewed as performance errors, however, the correct picture of human cognition seems to be that humans have a cognitive competence that includes heuristics which do not conform to norms of reasoning.[24]

NOTES

[1] Two classic experiments which are supposed to demonstrate human irrationality are the selection task that shows humans systematically violate rules of logic – see Wason

(1966, 1968) – and the conjunction experiment that shows humans systematically violate rules of probability – see Tversky and Kahneman (1983). More recently, some psychologists have argued that these experiments which are supposed to demonstrate irrationality are at best misleading. See, for example, Gigerenzer (1991a, 1991b).

[2] In addition to the arguments considered in this essay, other arguments for the claim that humans are rational include those made by Dennett (1987), Davidson (1984), Fodor (1981), Goldman (1986), Henle (1962, 1978), Lycan (1988), Millikan (1984), Papineau (1987), Popper (1984), Quine (1969), and Sober (1981). Some of these arguments are discussed by Stich (1990) and Stein (1991).

[3] Macnamara (1986) uses the term "mental logic", while Stich (1990) uses the term "psycho-logic", for what I, following Cohen (1981), call "cognitive competence".

[4] See, for example, Chomsky (1986, 1988).

[5] Cohen (1981); Macnamara (1986); and Sober (1978).

[6] Garfield (1988) briefly points to this difference between linguistics and reasoning.

[7] In the remainder of this paragraph, I paraphrase and quote from Rawls (1971, pp. 20–21), omitting reference to Rawls's notion of the "original position". Parenthetical references in the paragraph are to Rawls (1971).

[8] For a discussion of experiments that involve violation of principles of logic, see Wason and Johnson-Laird (1972, Chaps. 13–15) and Manktelow and Over (1990, Chap. 6). On experiments that involve violation of principles of probability, see Tversky and Kahneman (1983).

[9] The example comes frorn Stich and Nisbett (1980, pp. 191–93), as well as Stich (1990, pp. 83–84). Cohen (1986, pp. 169–73), however, argues that the gambler's fallacy is not a fallacy after all. For other examples of principles that would be in reflective equilibrium but do not seem rational, see Stich and Nisbett (1980, pp. 193–95).

[10] Stich and Nisbett (1980, pp. 197–98) refers to this response as "digging in".

[11] Stich (1990, pp. 83–86) discusses the first two views; presumably, he thinks that what he says against the first two views counts against the third.

[12] For further criticism of this modification to the expert view, see Conee and Feldman (1983) and Stich (1990, p. 164, n. 16).

[13] Cohen (1981, pp. 320, 323) explicitly rejects this suggestion. He says that the norms of reasoning "require a *narrow*, not a *wide*, reflective equilibrium" (p. 320, emphasis added).

[14] The distinction is implicit in Rawls (1971, p. 49).

[15] For discussion of the distinction between wide and narrow reflective equilibrium, see Daniels (1979, 1980a, 1980b).

[16] Stich (1990, Chap. 4) calls the project of attempting to use some variant of reflective equilibrium to justify norms of reasoning "the Neo-Goodmanian project". He argues that this project is doomed to fail. While examining this argument is beyond the scope of this paper, my general worry is that Stich dramatically underestimates the resources available to friends of reflective equilibrium.

[17] This is roughly how Jose R. Capablanca, a great chess master, allegedly learned how to play chess. According to Capablanca (1965), when he was five years old, he watched his father and a friend, both chess novices, play the game several nights in a row. After his father won a game by making an illegal move, the young Capablanca pointed out his father's error and proceeded to demonstrate his secretly learned ability to play chess.

[18] The emphasis on considered intuitions is present in Macnamara (1986, pp. 22–42), but without any talk of reflective equilibrium.

[19] For a discussion of introspection in the history and prehistory of cognitive science, see Flanagan (1984). For a detailed philosophical discussion of introspection in psychology, see Lyons (1986).

[20] Sternberg's results do not in fact indicate whether subjects "look" through the memorized lists from left to right, from right to left, or, as implausible as it may seem, in some other (perhaps random) order; what his data show is that subjects take the same amount of time to identify a number as being on the list if the number is at or near the beginning of the list as they do if the number is at or near the end.

[21] Cherniak's critique specifically applies to Quine's account of the heuristics needed to maintain the web of belief, not necessarily to other similar "network" theories of belief.

[22] Pinker and Bloom (1990) argue that our innate linguistic capacity is the result of natural selection. If they are correct, as I think they are, then perhaps evolutionary theory might be used to inform linguistic theory in the way that evolutionary theory can inform cognitive science. Pinker and Bloom's thesis is, however, highly contentious. For opposing views, see the commentaries on Pinker and Bloom (1990, pp. 727–65); also see Chomsky (1988), Gould (1987), Lewontin (1990), and Piattelli-Palmarini (1989).

[23] I consider other arguments for this claim in Stein (1991).

[24] Parts of this essay were read at the June, 1991 conference of the Society for Philosophy and Psychology in San Francisco. Thanks to Ned Block, Paul Bloom, David Brink, Tracy Isaacs, Diane Jeske, Gary Marcus, Paul Snowden, William Snyder, Bob Stalnaker, Daniel Stoljar, Mike Weber, Karen Wynn, and two anonymous referees for their very helpful comments and suggestions.

REFERENCES

Berwick, R., and A. Weinberg: 1980, *The Grammatical Basis of Linguistic Performance*, MIT Press, Cambridge.

Capablanca, J. R.: 1965, *My Chess Career*, Dover, New York.

Caplan, D., and N. Hildebrandt: 1988, *Disorders of Syntactic Comprehension*, MIT Press, Cambridge.

Cherniak, C.: 1986, *Minimal Rationality*, MIT Press, Cambridge.

Cherniak, C.: 1988, 'Undebuggability and Cognitive Science', *Communications of the Association for Computing Machinery* **31**, 402–12.

Chomsky, N.: 1986, *Knowledge of Language*, Praeger, New York.

Chomsky, N.: 1988, *Language and Problems of Knowledge*, MIT Press, Cambridge.

Cohen, L. J.: 1981, 'Can Human Irrationality Be Experimentally Demonstrated?', *Behavioral and Brain Sciences* **4**, 317–70.

Cohen, L. J.: 1986, *The Dialogue of Reason*, Oxford University Press, Oxford.

Conee, E., and R. Feldman: 1983, 'Stich and Nisbett on Justifying Inference Rules', *Philosophy of Science* **50**, 326–31.

Cosmides, L.: 1989, 'The Logic of Selection: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task', *Cognition* **31**, 1187–276.

Daniels, N.: 1979, 'Wide Reflective Equilibrium and Theory Acceptance in Ethics', *Journal of Philosophy* **76**, 256–82.

Daniels, N.: 1980a, 'On Some Methods of Ethics and Linguistics', *Philosophical Studies* **37**, 21–36.

Daniels, N.: 1980b, 'Wide Reflective Equilibrium and Archimedean Points', *Canadian Journal of Philosophy* **10**, 83–103.

Davidson, D.: 1984, *Inquiries into Truth and Interpretation*, Oxford University Press, Oxford.

Dennett, D. C.: 1987, *The Intentional Stance*, MIT Press, Cambridge.

Flanagan, O.: 1984, *The Science of the Mind*, MIT Press, Cambridge.

Fodor, J. A.: 1981, 'Three Cheers for Propositional Attitudes', in *Representations*, MIT Press, Cambridge, pp. 100–23.

Garfield, J.: 1988, 'Review of John Macnamara's *A Border Dispute*', *Journal of Symbolic Logic* **53**, 314–16.

Gigerenzer, G.: 1991a, 'How to Make Cognitive Illusions Disappear: Beyond "Heuristics and Biases"', *European Review of Social Psychology* **2**, 83–115.

Gigerenzer, G.: 1991b, 'On Cognitive Illusions and Rationality', in E. Eells and T. Maruszewski (eds.), *Probability and Rationality: Studies in L. Jonathan Cohen's Philosophy of Science*, Rodopi, Amsterdam, pp. 225–49.

Goldman, A.: 1986, *Epistemology and Cognition*, MIT Press, Cambridge.

Goodman, N.: 1983, *Fact, Fiction and Forecast*, 4th ed., Harvard University Press, Cambridge.

Gould, S. J.: 1987, 'The Limits of Adaptation: Is Language a Spandrel of the Human Brain?', talk presented to the Cognitive Science Seminar, Center for Cognitive Science, Massachusetts Institute of Technology, Cambridge.

Henle, M.: 1962, 'On the Relation Between Logic and Thinking', *Psychological Review* **69**, 376–82.

Henle, M.: 1978, 'Foreword', in R. Revlin and R. E. Mayer (ed.), *Human Reasoning*, Winston, Washington D.C., pp. xiii–xviii.

Lewontin, R. C.: 1990, 'The Evolution of Cognition', in D. Osherson and E. Smith (eds.), *Thinking: An Invitation to Cognitive Science*, Vol. 3, MIT Press, Cambridge, pp. 229–46.

Lycan, W.: 1988, 'Epistemic Value', in *Judgement and Justification*, Cambridge University Press, Cambridge, pp. 128–56.

Lyons, W.: 1986, *The Disappearance of Introspection*, MIT Press, Cambridge.

Macnamara, J.: 1986, *A Border Dispute*, MIT Press, Cambridge.

Manktelow, K. I., and D. E. Over: 1990, *Inference and Understanding: A Philosophical and Psychological Perspective*, Routledge, New York.

Millikan, R.: 1984, 'Naturalist Reflections on Knowledge', *Pacific Philosophical Quarterly* **65**, 315–34.

Papineau, D.: 1987, *Reality and Representation*, Basil Blackwell, Cambridge.

Parfit, D.: 1984, *Reasons and Persons*, Oxford University Press, Oxford.

Piattelli-Palmarini, M.: 1989, 'Evolution, Selection and Cognition', *Cognition* **31**, 1–44.

Pinker, S., and P. Bloom: 1990, 'Natural Language and Natural Selection', *Behavioral and Brain Sciences* **13**, 707–84.

Popper, K.: 1984, 'Evolutionary Epistemology', in J. W. Pollard (ed.), *Evolutionary Theory: Paths into the Future*, Wiley and Sons, London, pp. 239–56.

Putnam, H.: 1975, 'The Meaning of "Meaning"', in *Mind, Language and Reality: Philosophical Papers*, Vol. 2, Cambridge University Press, Cambridge, pp. 215–71.

Quine, W. V. O.: 1961, 'Two Dogmas of Empiricism', in *From a Logical Point of View*, Harvard University Press, Cambridge, pp. 20–46.

Quine, W. V. O.: 1969, 'Natural Kinds', in *Ontological Relativity and Other Essays*, Columbia University Press, New York, pp. 114–38.

Quine, W. V. O., and J. S. Ullian: 1970, *The Web of Belief*, Random House, New York.

Rawls, J.: 1971, *A Theory of Justice*, Harvard University Press, Cambridge.

Rawls, J.: 1974/1975, 'The Independence of Moral Theory', *Proceedings and Addresses of the American Philosophical Association* **48**, 5–22.

Sober, E.: 1978, 'Psychologism', *Journal of Social Behavior* **8**, 165–91.

Sober, E.: 1981, 'Evolution of Rationality', Synthese **46**, 95–120.

Stein, E.: 1991, *Rationality and the Limits of Cognitive Science*, Ph.D. dissertation, Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge.

Sternberg, S.: 1966, 'High-speed Scanning in Human Memory', *Science* **153**, 652–54.

Stich, S.: 1990, *Fragmentation of Reason*, MIT Press, Cambridge.

Stich, S., and R. Nisbett: 1980, 'Justification and the Psychology of Human Reasoning', *Philosophy of Science* **47**, 188–202.

Tversky, A., and D. Kahneman: 1983, 'Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement', *Psychological Review* **90**, 293–315.

Wason, P. C.: 1966, 'Reasoning', in B. Foss (ed.), *New Horizons in Psychology*, Penguin, Middlesex, pp. 135–51.

Wason, P. C.: 1968, 'Reasoning about a Rule', *Quarterly Journal of Experimental Psychology* **20**, 273–81.

Wason, P. C., and P. N. Johnson-Laird: 1972, *Psychology of Reasoning: Structure and Content*, Harvard University Press, Cambridge.

Wexler, K., and P. Culicover: 1980, *Formal Principles of Language Acquisition*, MIT Press, Cambridge.

Department of Philosophy
New York University
New York, NY 10003-6688
U.S.A.