RATIONALITY AND THE LIMITS OF COGNITIVE SCIENCE

by

EDWARD D. STEIN

B.A., Philosophy, Williams College (1987)

Submitted to the Department of Linguistics and Philosophy
in Partial Fulfillment of the Requirements for the
Degree of

DOCTOR OF PHILOSOPHY
in Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1992

Signature of author . . . .
Department of Linguistics and Philosophy
November 20, 1991

Certified by . . . .
Ned Block
Professor of Linguistics and Philosophy
Thesis Supervisor

Accepted by . . . .
Wayne O'Neil
Professor of Linguistics and Philosophy
Department Chairperson

For my parents, my grandparents

and the rest of my extended family

RATIONALITY AND THE LIMITS OF COGNITIVE SCIENCE

by Edward Stein

ABSTRACT: The observation that humans are often irrational has become commonplace. This observation has received empirical support from various experiments performed by cognitive scientists that are supposed to show that humans systematically violate principles of probability, rules of logic, and other norms of reasoning. In response to these experiments, philosophers have made creative and appealing arguments that these experiments must be mistaken or misinterpreted because humans must be rational. I examine these arguments for human rationality and show that they fail; cognitive science does play a role in assessing human rationality.

In Chapter One, I examine three cognitive science experiments in detail and discuss what is at issue in the debate about human rationality. Given that everyone agrees humans make mistakes in reasoning, at issue is whether the mistakes humans make are *mere* mistakes or indicative of systematic divergences from norms of reasoning. Defenders of human rationality say these mistakes are not characteristic of underlying reasoning abilities but are *performance* errors. Because humans make performance errors, human reasoning ability cannot be read off of behavior, but rather is to be studied by exploring *cognitive competence*, the underlying ability to reason.

In Chapter Two, I develop and discuss the notion of cognitive competence. I then consider and reject two arguments that human cognitive competence must match the norms of reasoning. The first argument appeals to simplicity considerations; the second to the supposed common origin of our cognitive competence and our normative standards from our intuitions about what counts as good reasoning. In Chapter Three, I turn to a related argument for the view that humans are rational that says cognitive competence is discovered in the same way norms of reasoning are justified: both projects involve *reflective equilibrium* using as data people's judgments about what counts as good reasoning. Human cognitive competence thus matches the normative standards; humans are thereby rational. I argue that even if reflective equilibrium is the right model for studying cognitive competence and justifying norms of reasoning, there are strong reasons to think that the results of these two processes will diverge, in other words, that humans might still be irrational.

In Chapter Four, I consider an argument for human rationality that draws on the principle of charity: when translating utterances of a speaker, one should try to interpret her utterances as rational and true. Applied to reasoning, this principle advocates characterizing human cognition as rational. This counts against interpreting psychological experiments as demonstrating human irrationality. I argue that a *strong* notion of the principle of charity is required to support the claim that humans must be rational, but only a *weak* version of this principle is defensible with respect to the interpretation of intentional states.

In Chapter Five, I consider an argument for human rationality that appeals to evolution. This argument claims (i) that natural selection will select for cognitive mechanisms that produce true beliefs, and (ii) that such "truth-tropic" mechanisms are sufficient for being rational. I argue that a common criticism of this argument—that various non-selectionist forces are responsible for the evolution of cognitive mechanisms—fails. Even so, the evolutionary argument, even in its most sophisticated forms, also fails because the filter of natural selection is too coarse-grained to select only the truth.

My overall conclusion is that while cognitive science does not determine what the normative principles of reasoning are, it is relevant to whether or not human cognitive competence matches these principles; whether or not humans are irrational is, thus, an empirical question.

ACKNOWLEDGMENTS

For me, the writing of a philosophy dissertation has, in a variety of ways, been a very different project from writing a journal-length philosophy paper. For instance, the number of times per page the writer of a dissertation can say "this question is beyond the scope of this project" is greatly reduced compared to a paper. Also, the writer of a dissertation is forced to keep a significantly larger set of questions in his or her head at the same time. Another difference in writing a dissertation has been the sort of help that I could get from friends and colleagues. Some philosophers find that they write best when they work like hermits. In most of my previous work, I found this not to be the case. I work best when I am discussing my ideas with others who know my topic well, with others who know the general area I am working on but not my particular topic, as well as with those who are intelligent but naive interlocutors. However, with respect to this project, I have worked more like a hermit than ever before. This has meant that I have been more dependent on my faculty advisors and on friends and colleagues who I can persuade to read drafts of what I have written compared to other projects on which I was able to get advice and motivation through less structured interaction (for example, over dinner). Also, because of the breadth of my dissertation, I have been more dependent than usual on advice from others on topics outside my areas of specializations, for example, ethics and linguistics.

Despite my relative solitude, there have been dozens of people who have helped me in various ways with this project over the past few years. The following people provided general help with the project, in most cases by being provocative and intelligent conversation partners: Michael Antony, Alyssa Bernstein, Ned Block, Paul Bloom, David Brink, Ron Caldwell, Susan Dwyer, Steve Greene, Tracy Isaacs, Diane Jeske, Tom Kuhn, Peter Lipton, Eric Lormand, Gary Marcus, Paul Pietrowski, Georges Rey, Scott Smith, Paul Snowden, William Snyder, Bob Stalnaker, Daniel Stoljar, Mike Weber, Rob Wilson, and Karen Wynn. Theresa Tobin of the Humanities Library at M.I.T. and Patricia Claffey of the Brain and Cognitive Science Library at M.I.T. provided friendly and excellent research assistance. The following people provided specific help on the noted chapters:

Chapter One: Paul Bloom, David Brink, Steve Greene, Eric Lormand, Gary Marcus, William Snyder, Karen Wynn

Chapter Two: Paul Bloom, Ned Block, David Brink, Gary Marcus, William Snyder, Bob Stalnaker

Chapter Three: Paul Bloom, Ned Block, David Brink, Diane Jeske, Gary Marcus, Paul Snowden, William Snyder, Bob Stalnaker, Mike Weber, Karen Wynn

Chapter Four: Ned Block, Paul Bloom, David Brink, Bob Stalnaker

Chapter Five: Paul Bloom, Ned Block, David Brink, Leda Cosmides, Peter Lipton, Eric Lormand, Paul Pietrowski, Bob Stalnaker

Chapter Six: Ned Block, Bob Stalnaker

My advisors in particular need to be singled out for their efforts. Paul Bloom, although he was not formally a member of my advisory committee behaved, to my pleasure and benefit, as if he was. From the earliest days of this project to the bitter end, he promptly and carefully read every bit of writing I showed him (and that included quite a lot) and made thorough, provocative and wide-ranging comments. His grasp of a very wide range of important linguistic, psychological and philosophical issues continually impressed and assisted me. He never gave up trying to get me to understand ideas that I sometimes stubbornly refused to fully appreciate. Without his comments, questioning and advice the process of writing would have been less fun and less educational and the end-product of this process would have been noticeably weaker.

The (formal) members of my committee were all active participants in this project; each of their individual contributions were significant and well-complemented those of their colleagues. David Brink was an especially careful reader of my drafts. His comments were always detailed, precise, and sympathetic. He also helped me to see connections with issues that seemed only superficially connected to my central worries, but that ultimately proved quite relevant and philosophically rich. These connections produced a dissertation that is connected to more philosophical issues than I originally expected. Also, particularly

early on in the project, David helped me to grasp the scope and limits of my project. Ned Block, the chairman of my committee, was an enthusiastic supporter throughout this project. He raised an assortment of helpful and interesting objections, and always offered frank and direct criticism. Every meeting with Ned produced at least one major insight that soon found its way into the pages of this project. Finally, Bob Stalnaker was always quick to read my drafts, generous with his time, and immensely provocative in discussion. He continually encouraged me to see the big picture, pushing me on the connections among the various issues that concerned me. His calm, systematic, friendly method of questioning acted as a catalyst by being both a source of encouragement and of puzzlement. His suggestions richly improved this dissertation. Without the comments and assistance of Paul, David, Ned and Bob, this manuscript would have been, at best, a shadow of its current form.

The work carried out on this project was done while I was a graduate student at M.I.T. For much of this time, I received financial support in the form of tuition waivers and teaching or research stipends from M.I.T.'s Department of Linguistics and Philosophy. I also received a year of dissertation support from the Woodrow Wilson Foundation through the Mellon Fellowship in the Humanities. Additionally, I spent one semester of the time working on this project as a visiting instructor in the Department of Philosophy at Williams College. To these institution and many of the individuals involved with them, I am deeply grateful.

I have many other people to thank for their assistance who are not even mentioned above. This list would include my roommates over the past few years (Eric, Tina, Brian, Pam and Lori), my "boyfriends" over the past few years (Dan, Tom and Raymond), members of my extended family, and friends too numerous to mention them all. Despite this help and support, I alone deserve the credit for any (performance or competence) errors that may be contained in what follows.

CONTENTS

I wonder who it was defined man as a rational animal.
It was the most premature definition ever given.

Oscar Wilde, *The Picture of Dorian Gray*

Chapter One: Introduction

I.

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 reasoning capacities has garnered scientific support in the last quarter century from 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 that 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. These arguments go to the heart of theory of knowledge, as well as philosophy of science, mind and language. My project here is to develop and evaluate these arguments in order to determine whether and in what sense humans are rational.

The claim that humans are irrational (what I call the irrationality thesis) is 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. A person who systematically, for example, believes both that p is true and that q follows from p but does not believe q 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

[1] Among those who have claimed that such experiments support the conclusion that humans are irrational include: Daniel Kahneman and Amos Tversky, "Subjective Probability: A Judgement of Representativeness," *Cognitive Psychology* 3 (1972), reprinted in *Judgement under Uncertainty· Heuristics and Biases*, Daniel Kahneman, Paul Slovic and Amos Tversky, eds. (Cambridge: Cambridge University Press, 1982), p. 46 (in *Judgement* anthology), "[F]or anyone who would wish to view man as a reasonable intuitive statistician, such results are discouraging"; R. E. Nisbett and E. Borgida, "Attribution and the Psychology of Prediction," *Journal of Personal and Social Psychology* 32 (1975), p. 935, "[these experiments] have bleak implications for human rationality"; Paul Slovic, Baruch Fischhoff and Sarah Lichtenstein, "Cognitive Processes and Societal Risk Taking," in *Cognition and Social Behavior*, J. S. Carroll and J. W. Payne, eds. (Hillsdale, NJ: Ersbaum, 1976), p. 173-174, "people's judgments of important probabilistic phenomena are not merely biased but are in violation of fundamental normative rules"; and Richard Nisbett, David Kranz, Christopher Jepson, and Ziva Kunda, "The Use of Statistics in Everyday Inductive Reasoning," *Psychological Review* 90 (1983), p. 340, "people commit serious errors of inference."

thesis (what I call the irrationality experiments) are said to provide evidence for such claims.

Those who defend the claim that humans are rational (what I call the rationality thesis) need to undermine the irrationality experiments. There are two interconnected ways that they try to do this. First, they argue that the irrationality experiments are simply being misinterpreted when they are taken as support for the irrationality thesis. Empirical evidence may be helpful in supporting this approach to defending the rationality thesis. Second, they develop general conceptual arguments that humans are rational. These arguments entail that cognitive science and psychology are constrained in such a way that they cannot establish human irrationality. On this view, cognitive science can no more demonstrate that humans are irrational than it can demonstrate that humans are moral or that humans have good aesthetic sensibilities. Some see this limitation on cognitive science as a limitation on *any* empirical inquiry while others see it as applying only to cognitive science either because of epistemological constraints on the field or because whether humans are rational comes under the purview of some other discipline (for example, biology). Most of the arguments for the rationality thesis that I will be considering are of the conceptual sort, but empirical considerations will come into play at various times as well.

In the rest of this chapter, I will describe the lay of the land around the irrationality thesis, saying where and how various issues fit together. I will begin with a detailed discussion of the cognitive science experiments that are claimed to count as evidence for the view. Next, I will say more about the two theses under considerations: some versions of these theses about human rationality are uncontentious; I will focus on a contentious version of them with an eye towards specifying what the core issue is between the two theses. I will then map out the various arguments for and against the irrationality thesis that I will be considering in subsequent chapters. Finally, I will say something about what these arguments have to do with the limits of cognitive science.

II.

To begin, I will discuss three cognitive science experiments that seem to bear on human rationality. These particular experiments have been chosen both because they are discussed by philosophers and because there is an extensive psychological literature on them. What matters, however, are not so much the particular details of the experiments, but the general features of them that can be construed as supporting the irrationality thesis; these experiments are chosen to be representative of those cognitive scientists appeal to when they claim that humans are irrational. I will look first at the Wason selection task, an experiment that is supposed to show that humans make systematic logical errors; second, at the conjunction experiment, an experiment that is supposed to show that humans ignore basic rules of probability; and, third, at an experiment that is supposed to show that humans often take irrelevant data into consideration.

In 1966, P. C. Wason published the results of an experiment that tested human ability to apply principles of logic.[2] In this experiment, subjects are presented with four cards, each with a number on one side and a letter on another. The four cards are showing a vowel, a consonant, an even number and an odd number, respectively, say 'A,' 'K,' '4,' and '7.' Subjects are then asked which cards they will need to turn over to test the truth of the rule "if a card has a vowel on one side, then it has an even number on the other side." Most of the subjects answered that either the cards showing 'A' and '4' need to be turned over to test the rule or just the card showing 'A' needs to be turned over. The correct solution is to turn over the 'A' and the '7' card; very few subjects said the '7' card should be one of the cards that is turned over.

[2] The pilot study is P. C. Wason, "Reasoning," in *New Horizons in Psychology*, B. Foss, ed. (Middlesex, England: Penguin, 1966), pp. 135-151. A more extensive experiment is reported in Wason, "Reasoning about a Rule," *Quarterly Journal of Experimental Psychology* 20 (1968), pp. 273-281. For a detailed discussion, see Wason and P. N. Johnson-Laird, *Psychology of Reasoning: Structure and Content* (Cambridge: Harvard University Press, 1972), chapters 13-15. For a recent summary, see K. I. Manktelow and D. E. Over, *Inference and Understanding: A Philosophical and Psychological Perspective* (New York: Routledge, 1990), Chapter Six.

table 1: truth table for "if p, then q"

p	q	if p, then q
T	T	T
T	F	F
F	T	T
F	T	T

That the 'A' and '7' cards are the ones that should be turned over is true by the straightforward application of logic. The rule subjects are asked to test ("if a card has a vowel on one side, then it has an even number on the other side") is a simple conditional statement of the form "if p, then q." For a conditional to be true, it must be the case that when the antecedent ("the card has a vowel on one side," that is, p) is true, the consequent ("the card has an even number on the other side," that is, q) must be true (see table 1 for a truth table of "if p, then q"). To test a conditional, one looks for cases in which the antecedent (p) is true but the consequent (q) is false. Thus, the 'A' card needs to be checked because if it has an odd number on its back, the antecedent will be true but the consequent false. Looking at the first row of table 2, the truth value of the conditional "if p, then q" can*not* be assessed without knowing whether the other side of the 'A' card has an odd or even number on it (that is, whether q is true or false). The 'K' card need not be checked because the antecedent is false—no matter what is on the other side (an odd or even number), the conditional will be true. Looking at table 2, the truth value of the conditional "if p, then q" in the 'K'-card case *can* be known without looking at the unseen side of the card; if the antecedent is false, then the conditional is true no matter what the truth value of the antecedent is. The '4' card, for similar reasons, need not be flipped over to test the truth of the conditional. Whether there is a vowel or consonant on the other side, the rule in question will not be violated. The conditional in the '4'-card case will be true whether or not tne antecedent is. Finally, the '7' card *does* need to be examined to see whether it has a vowel on the back; if it does, then the conditional is false (because the antecedent would be true—the card has a vowel on it—and the consequent would be false—the card does not have an even number on it) and if it does not, then the conditional is true (because the antecedent would be false).

table 2: truth table for the selection task

	p	q	if p, then q
case 'A'	T	?	?
case 'K'	F	?	T
case '4'	?	T	T
case '7'	?	F	?

As this discussion has shown, subjects should have examined the 'A' and '7' cards; most of them, however, failed to do so. This error seems to be well-entrenched (that is. individuals will repeat the error) and common.[3] Studies have shown, however, that in slightly different contexts, subjects do much better at applying the very same rules that they fail to properly apply in the standard selection task. Consider a concrete version of the selection task in which subjects are presented with a deck of cards each with the name of a city on one side of it and a mode of transportation on the other. Subjects are told to imagine each card as representing a trip made by the experimenter. They are asked to test whether a rule like "Eve:y time I go to New York, I travel by train" is true of four cards respectively showing 'New York,' 'Philadelphia,' 'train,' and 'car.' Almost two-thirds of the subjects correctly select the cards showing 'New York' and 'car' to turn over; this is

[3] For evidence that it is well-entrenched, see P. C. Wason, "Regression in Reasoning?" *British Journal of Psychology* 60 (1969), pp. 471-480; also see Wason and P. N. Johnson-Laird, "A Conflict Between Selecting and Evaluating Information in an Inferential Task," *British Journal of Psychology* 61 (1970), pp. 509-515.

dramatically better than the less than ten percent who picked the correct cards in the standard selection task.[4]

Various explanations for the improved performance on the concrete selection task as compared to the abstract selection task have been offered including the idea that the *realism* of the concrete task facilitates the translation of the selection task into a form with which subjects could better reason[5] and the idea that the *familiarity* of the concrete task is the crucial difference between them.[6] Neither of these ideas, however, have proved adequate.[7] More recently, the favored strategy for explaining the varied response that subjects give to the different versions of the selection task is that reasoning is content-dependent, that is, whether subjects invoke the right rules of reasoning depends on the *content* of the particular version of the task. Different psychologists have developed accounts of how this content-dependency effect works. Two of the most interesting accounts are those of Cheng and Cosmides. Cheng's account has to do with whether the particular version of the experimental task fits the pattern involved in the granting of permission (for example, "If the person is driving a car, then she must be sixteen years or older")[8] while Cosmides's account has to do with whether the version of the task fits the pattern involved in social exchange situations (for example, "If you help me plow my fields, then I will help you milk your cows").[9] Briefly, Cosmides's idea is that people will do well on versions of the

[4] The experiment discussed is from P. C. Wason and D. Shapiro, "Natural and Contrived Experience in a Reasoning Problem," *Quarterly Journal of Experimental Psychology* 23 (1971), pp. 63-71. P. N. Johnson-Laird, P. Legrenzi and M. S. Legrenzi, "Reasoning and a Sense of Reality," *British Journal of Psychology* 63 (1972), pp. 395-400, discusses another concrete version of the selection task. These experiments and other related ones are discussed in *Psychology of Reasoning*, chapters 14 and 15.

[5] For example, "Reasoning and a Sense of Reality."

[6] R. A. Griggs, "The Role of Problem Content in the Wason Selection Task and THOG Problem," in *Thinking and Reasoning: Psychological Approaches*, J. Evans, ed. (London: Routledge and Kegan Paul, 1983).

[7] For a critique of the realism idea, see for example K. I. Manktelow and J. Evans, "Facilitation of Reasoning by Realism: Effect or Non-Effect," *British Journal of Psychology* 70 (1979), pp. 477-488. For a critique of the familiarity effect, see for example P. N. Johnson-Laird, *Mental Models* (Cambridge: Cambridge University Press, 1983).

[8] P. W. Cheng and K. J. Holyoak, "Pragmatic Reasoning Schemas," *Cognitive Psychology* 17 (1985), pp. 391-416.

[9] Leda Cosmides, "The Logic of Selection: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task," *Cognition* 31 (1989), pp. 187-276

selection task that fit the form of social exchange situations and that involve the detection of "cheaters" with respect to the social contract (that is, those who try to get the benefit without paying the cost for doing so, for example, I would be a cheater if I did not help you milk your cows after you helped me plow my fields). This explanation is supposed to account for the fact that subjects reason well on certain concrete tas^k^s but not on others.[10] The crucial point for the moment is that whatevei the psychological explanation for the varied responses to the selection tasks, subjects' responses in the standard abstract selection task (henceforth, unless otherwise noted, I shall call it simply the selection task) as well as certain versions of the concrete selection task, seem, at least at first blush, irrational—when it comes to inferences involving deductive logic, people seem to diverge from the norms. This seems to provide good reason for thinking humans are irrational.

The second experiment, published in 1983 by Amos Tversky and Daniel Kahneman,[11] is claimed to show that humans, in making judgments about the likelihood of various events, systematically violate a basic and important principle of probability. Subjects read a description of a person like the following:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nucleai demonstrations.[12]

Subjects are then asked to rate the likelihood of various statements about this person (in this case, Linda's current profession and political affiliation). Among the options they are asked to rank are the following:

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

[10] Cosmides's idea is particularly interesting because she claims to draw support for it from evolutionary theory. I will return to this later. Cosmides's proposal has been criticized by P. W. Cheng and K. J. Holyoak, "On the Natural Selection of Reasoning Theories," *Cognition* 33 (December 1989), pp. 285-333; and Paul Pollard, "Natural Selection for the Selection Task: Limits to Social Exchange Theory," *Cognitinn* 36 (August 1990), pp. 195-204.

[11] Amos Tversky and Daniel Kahneman, "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement," *Psychological Review* 90 (October 1983), pp. 293-315.

[12] *Ibid.*, p. 297.

Eighty-five percent of the subjects (including both statistically naive undergraduates and presumably *not* statistically naive psychology graduate students) ranked (3) as being more probable than (2). In so doing, they made a serious mistake. It is a basic rule of probability (called the conjunction rule) that the probability of two events occurring together must be less than or equal to the probability of either one of the two events occurring individually. This is because any instance of both of the events occurring is *necessarily* an instance of a particular one of them occurring while every instance of a particular one of them occurring is *not* necessarily an instance of both events occurring. In other words, it is possible that one event might occur without both events occurring, but it is *not* possible that both events might occur without either one of them also occurring. In particular, Linda could be a bank teller but *not* a feminist bank teller (for instance, her career in banking might have turned her attention away from issues of social justice), but she could *not* (under any scenario) be a bank teller and a feminist yet *not* also a feminist. Subjects who say Linda is more likely to be a bank teller *and* a feminist than she is to be a bank teller violate the conjunction rule.

Violating the conjunction rule is a serious mistake. If a person consistently bets her money in violation of this rule, then she would lose all her money. In more precise terms, a person who violates a rule of probability theory is susceptible to what is known as a Dutch book. A person is "Dutch-book-able" if there is a series of bets that she will accept such that, no matter what the outcome of the events bet on, she will lose all of her money.[13]

Suppose someone (call him Henry) is willing to bet in accordance with the likelihood judgments subjects make about Linda, that is, he is more willing to bet that Linda is a bank teller *and* a feminist than he is willing to bet that she is a bank teller. If this is the case, Henry can be Dutch-booked. For example, suppose Henry thinks the probability that

[13] For a general brief discussion of Dutch books, see Brian Skyrms, *Choice and Chance* (Belmont, CA: Wadsworth, 1986), pp. 185-189.

Linda is a feminist is .5, the probability that Linda is a bank teller is .1 and that Linda is a feminist *and* a bank teller is .25. (Note that these probabilities match with the responses the majority of subjects gave in the conjunction experiment; Henry rates the probability that Linda is a bank teller *and* a feminist as greater than the probability that Linda is a bank teller.) Suppose further that a bookie offers Henry the following three bets:

(a) A one dollar bet at 1-to-1 odds that Linda is *not* a feminist.
(b) A six dollar bet at 1-to-9 odds that Linda is *not* a bank teller.
(c) A two dollar bet at 3-to-1 odds that Linda is both a feminist and a bank teller.

Given Henry's probability ratings, these are all fair bets, that is, they are bets that Henry should accept because he has a fair chance of winning.[14] If he accepts all of these bets, whatever turns out to be true of Linda, Henry will lose money. To show this, here are four cases to examine, one for each possible outcome:

(i) Linda is both a bank teller and a feminist
(ii) Linda is a bank teller but not a feminist.
(iii) Linda is a feminist but not a bank teller.
(iv) Linda is neither a feminist nor a bank teller.

If Linda is both a bank teller and a feminist, Henry will win six dollars from bet (c) but lose six dollars from bet (b) and one dollar from bet (a) for a total loss of one dollar. If Linda is a bank teller but not a feminist, Henry will win one dollar from bet (a) but lose six dollars from bet (b) and two dollars from bet (c) for a total loss of seven dollars. If Linda is a feminist but not a bank teller, Henry will win sixty-six cents from bet (b) but lose one dollar from bet (a) and two dollars from bet (c) for a total loss of two dollars and thirty-four cents. Finally, if Linda is neither a bank teller nor a feminist, Henry will win one dollar from bet (a) and sixty-six cents from bet (b) but lose two dollars from bet (c) for a total loss of thirty-four cents. No matter what happens, Henry will lose when he bets according to probability ratings that violate the conjunction rule. Betting money when there is no chance

[14] A bet is fair if the expected value of a bet is equal to zero. The expected value of a bet is the sum of the products of the payoff of each result of the event bet on and the probability of that result occurring. So, bet (a) is fair because the probability that Linda is a not a feminist (.5) multiplied by the payoff (1) minus the probability that Linda is a feminist (.5) multiplying by the payoff (-1) equals zero. Similarly, for bet (b), .9 multiplied by 1/9 minus .1 multiplied by -1 equals zero and, for bet (c), .25 multiplied by 3 minus .75 multiplied by -1 equals zero. For a discussion of fair bets and expected value, see *Choice and Chance*, chapters 5 and 6.

of winning, which is what people who violate the conjunction rule are committing themselves to doing, is clearly irrational behavior.

The third experiment shows that humans take obviously irrelevant information into consideration when making predictions and decisions.[15] Subjects are asked to estimate various percentages like the percentage of African nations that are members of the United Nations. To begin, a roulette-like wheel is spun in a subject's presence to select a number between zero and one hundred. The subject is first asked to indicate whether the number that came up on the wheel is higher or lower than the percentage of African nations in the U. N. and then asked to estimate this percentage. Subjects who receive lower numbers on the roulette wheel consistently guess much lower than subjects who receive higher numbers on the wheel. For example, people who receive a spin of ten give a median estimate of twenty-five percent while people who receive a spin of sixty-five give a median estimate of forty-five. But clearly, the result of a random spin of a roulette wheel is totally *irrelevant* to the percentage of African nations in the United Nations. It seems irrational to factor, as humans seem to do, such irrelevant data into one's decision making.

In all three of these experiments, subjects behave in ways that seem to violate well-establishea norms of reasoning. Other experiments seem to demonstrate similar divergence between actual practice and established norms in other parts of reasoning.[16] Examples need not, however, be multiplied at this point; the primary question to address is what should be made of the results of these experiments, that is, should they be interpreted as showing that humans are irrational? A prior question, however, is what sorts of reasoning behaviors count as irrational. I will address this question in the next section.

[15] Paul Slovic and Sarah Lichtenstein, "Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgement," *Organizational Behavior and Human Performance* 6 (1971), pp. 649-744. The experiment discussed was done primarily to study the phenomenon called anchoring, the making of insufficient adjustments to estimates because of the point from which estimates started, namely guesses are skewed towards the starting point. See Tversky and Kahneman, "Judgement under Uncertainty: Heuristics and Biases," *Science* 185 (1974), pp. 1124-1131, reprinted in *Judgement under Uncertainty*, pp. 3-20; for discussion of anchoring, see especially pp. 14-18.

[16] See, for example, the essays in *Judgement under Uncertainty*.

III.

Having discussed these three experiments, I return to the two theses I introduced in the first section—the rationality thesis and the irrationality thesis. While it should be clear that these theses differ over the issue of whether or not humans are rational, I have not said what exactly the claim that humans are (or are not) rational amounts to. I turn now to this question.

No one would deny that humans are rational in the sense that our reasoning usually sustains us. The infeiences we make (for example about what food will nourish us) usually result in beliefs that are (at least) adequate in helping us continue to exist. Similarly, no one would deny that humans are irrational in the sense that humans often make mistakes in reasoning. Some of the inferences we make are due to cognitive slips, forgetfulness, and the like. If 'rational' is supposed to suggest complete freedom from reasoning errors, then humans are clearly *not* rational. The sense of rational that I am concerned with and that is invoked by the rationality and irrationality theses must be such that neither thesis is obviously true or false. For example, for this reason, Jonathan Bennett's definir'on of rationality as the sort of reasoning that is unique to humans cannot be the sense of rationality involved in the competing theses. Bennett writes, "I use 'rationality' to mean 'whatever it is that humans possess that marks them off, in respect of intellectual capacity, sharply and importantly, from all other known species.'"[17] With Bennett's definition of what it is to be rational inserted into the rationality thesis, the thesis is true by definition.[18] But this is not what friends of either the rationality or the irrationality thesis have in mind by 'rational.' Two other senses of rationality that would produce *un*interesting versions of the rationality and irrationality theses are two of the three

[17] Jonathan Bennett, *Rationality* (London: Routledge and Kegan Paul, 1964), p. 5.

[18] Bennett, *ibid.*, admits this. He says that it follows from his definition of rationality that "humans are the only rational creatures . . . , though it does not follow from the definition that it is useful or interesting to describe humans as 'rational.'"

senses of rationality that Jarvie and Agassi distinguish among.[19] The weakest sense Jarvie and Agassi discuss is that an agent is rational if it reasons in a goal-directed manner, the middle-strength sense is that an agent is rational if it reasons so as to obey explicit rules, and the strongest sense is that an agent is rational if it reasons in accordance with the highest standards of reasoning. No one is likely to deny that humans are rational in the two weakest senses Jarvie and Agassi propose, and, further, the irrationality experiments do nothing to shed doubt on whether humans are rational in either of these two senses. Only the third sense of rationality is called into question by the irrationality experiments; it is this sense of rationality that the defenders of the irrationality thesis deny humans have. Roughly, the idea of the rationality thesis is that there are normative standards of reasoning and that human reasoning conforms to these standards.

This idea of what it is to be rational is uninformative without saying what are the normative standards to which, according to the rationality thesis, humans conform and which, according to the irration ılity thesis, humans fail to conform. These normative standards include rules of logic (for example, if I believe p and I believe p implies q, then, according to the *modus ponens* norm, I ought to believe q), principles of probability (for example, according to the conjunction rule norm in probability theory, I ought to believe that the probability of p is greater than or equal to the probability of p and q), and the like. Developing a complete list of these normative standards is beyond the scope of this project; suffice to say that the forementioned principles and others like them would be on such a list.

But the further question remains: what is it to conform to normative standards? A tempting quick answer is that to conform to the norms of reasoning is to always do what the norms would dictate, for example, to always believe q when you believe p and p implies q, to always believe that the probability of p is greater than the probability of p and

[19] I. C. Jarvie and J. Agassi "The Problem of the Rationality of Magic," in *Rationality*, Brian Wilson, ed. (New York: Harper and Row, 1970), pp. 172-193; and "The Rationality of Dogmatism," in *Rationality Today*, T. Geraets, ed. (Ottowa: University of Ottowa Press, 1979), pp. 353-362.

q, and so on. This quick answer cannot be what friends of the rationality thesis have in mind for two important reasons.

First, always doing what the norms of reasoning dictate cannot be what we mean by rational because doing so is physically impossible given what Christopher Cherniak has called the "finitary predicament" of humans.[20] The finitary predicament is that (individual) humans have only a limited amount of time, a limited amount of memory capacity, a limited amount of processing power, and so on, available for reasoning. Given these constraints, it is not possible for humans to reason according to certain norms. Consider, for example, the norm of logical consistency: before coming to believe p, a person should check to see that she does not already believe not-p or something logically equivalent to it; if so, she must either not come to believe p or give up the belief with which p is inconsistent. Checking p against every other belief (and everything logically equivalent to them) would take too much processing time, so much so that always following the norm of logical consistency would be suicidal since it would require that an agent devote all of her time to checking for consistency, thereby neglecting to, for example, devote time to finding food to eat.[21] It seems wrong to call me irrational for failing to devote my time to finding out whether my new belief p is compatible with all my other beliefs when doing so would take an inordinate amount of time, time that would be better spent on other tasks.

Further, on reflection, always doing what the norms of reasoning dictate is clearly not a necessary condition for counting someone as rational. A person can make all sorts of mistakes in reasoning without being disqualified as rational. Suppose I know George is in his office, I believe that *if* George is in his office, then it must be Friday, and, further, suppose I have not gotten enough sleep last night. When you ask me what day it is and I say "Thursday," you do not, even if you know my beliefs about George and his office hours, accuse me of being irrational; rather, you accuse me of making a mistake, of being

[20] Christopher Cherniak, *Minimal Rationality* (Cambridge: MIT Press, 1986), p. 8.
[21] *Ibid.*, pp. 16-18.

forgetful, or the like, as a result of not having had enough sleep. You would remind me that I know George is in his office and that George is in his office only on Fridays, and then expect me to admit my mistake and confess that it must be Friday. Looking back on my behavior, you would *not* say that I was being irrational. If, however, I agree that George is in his office and that George is in his office only on Fridays, but continued to deny that it is Friday and did so without any attempt to reconcile my beliefs (for example, by saying that it is not always true that George is in his office only on Fridays, for instance, this week he is going to be in Paris on Friday so he came in on Thursday), *then* you might go so far as to call me irrational. But this suggests that more is involved in failing to be rational than just failing to follow norms of reasoning; sometimes, failing to follow norms is not a sign of irrationality, but rather is just due to having made some mistakes.

Behind the distinction between a mere mistake and a behavior actually due to irrationality is the idea that making a mistake involves a momentary lapse, a divergence from some typical behavior, while behavior due to irrationality connotes a divergence from the norm that is the result of something systematic. It is this distinction between mere mistakes and systematic violations of norms that friends of the irrationality thesis implicitly assume when they cite the irrationality experiments as evidence for their view. Consider, by way of example, the conjunction experiment. Subjects in this experiment are given a description of Linda and asked to say whether they think it is more likely that Linda is a bank teller or that Linda is both a bank teller and a feminist. The conjunction rule, a norm of reasoning, dictates that the probability that Linda is a bank teller is greater than or equal to the probability that Linda is both a bank teller and a feminist. If all humans on occasion violated this rule (by believing that it is more probable that Linda is a bank teller and a feminist than that she is a bank teller) but *usually* do not or if only a few humans regularly violate this rule but most do not, then these results would not count against the rationality thesis and these divergences from the normative standards could be interpreted as mere mistakes. But Kahneman and Tversky's experiments show that violation of the conjunction rule is frequent both for individuals and across humans in general. These results (at least *prima facie*) are supposed to count against the rationality thesis; this provides further reason for reading the irrationality thesis as claiming that humans systematically diverge from the norms of reasoning and for reading the rationality thesis as claiming that humans typically reason in accordance with these norms.

The distinction between mere mistakes and systematic errors underscores the point that what is at issue between the rationality thesis and the irrationality thesis has to do with *capacities*. I can have the capacity to do something (for example, ride a bike or apply *modus ponens*) and yet not display that capacity on a particular occasion (for example, because I am tired or drunk). The rationality thesis claims that humans have an underlying capacity to reason in accordance with the norms; the irrationality thesis denies this.

Talk of capacities leads nicely to another way of describing what is at issue between the rationality thesis and the irrationality thesis that involves borrowing the distinction between competence and performance from linguistics. A person's linguistic competence is her underlying knowledge of language, her ability to understand and utter grammatical sentences. People, however, 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. Rather they are due to some sort of interference with this competence, interference that prevents people from linguistic behavior in accordance with linguistic competence. Examples of these interferences include non-linguistic factors like insufficient memory, lack of attention, high amounts of alcohol in the blood stream, and so on. 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 ignore the static of performance errors that often affect actual linguistic behavior. Defenders of the rationality thesis see all of the divergences from the norms of reasoning that humans make as performance errors and, as such, they do not see these errors as indicative of human's underlying abilities 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 ou: "cognitive competence"[22] matches the norms of reasoning. For example, friends of the irrationality thesis would claim that the conjunction experiment shows that following the conjunction rule is *not* part of our cognitive competence; ın general, they would argue that our cognitive competence does not match the norms of reasoning and humans are thereby irrational. I will explore the analogy between linguistic competence and cognitive competence at length in Chapter Two; for now, I adopt the following terminology for discussing the two theses about human rationality: 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.

IV.

In the previous section, I used the fact that humans are in a finitary predicament as evidence that the rationality thesis could not mean that humans never diverge from the norms of reasoning. Some, however, have taken the finitary predicament as reason to construe the rationality thesis in a way quite different from what I have suggested so far; they suggest that because humans are in a finitary predicament, the norms of reasoning are too high a standard for humans to be held up against.[23] Instead, the question of whether humans are rational should be construed as asking whether humans are *as rational as they can be given the various constraints in the face of which they reason.* On this reading of what is at issue when we ask whether humans are rational (call this the *constrained* reading of rationality), the rationality thesis would be that humans are as rational as they can be

[22] John Macnamara, *A Border Dispute*, (Cambridge: MIT Press, 1986) uses the term "mental logic" while Stephen Stich, *Fragmentation of Reason* (Cambridge: MIT Press, 1990), uses the term "psycho-logic" for what I, following L. J. Cohen, "Can Human Irrationality Be Experimentally Demonstrated?," *Behavioral and Brain Sciences* 4 (1981), pp. 317-370, call cognitive competence.

[23] For example, *Minimal Rationality*.

given the constraints they face, while the irrationality thesis would be that humans are *not* as rational as they can be given these constraints.

Note that if this is the right way to see what is at issue between the rationality thesis and the irrationality thesis, then it is no longer obvious that the irrationality experiments are at all relevant to either the rationality or irrationality thesis. Consider the selection task as an example. *Prima facie*, the results of this experiment show that people systematically fail to apply the rules of logic when they are reasoning. On the reading of the rationality thesis discussed before this section (that is, the *un*constrained reading of the rationality thesis), these results are relevant because they seem to show that people diverge from the norms. But on the constrained reading of the rationality thesis, for the results to be relevant, one would have to show that it is possible, given the constraints humans face, for them to systematically make the right choices in the selection task. But showing this would be quite difficult. In fact, the finitary predicament and the irrationality experiments seem to count against this possibility. Thus, adapting the constrained reading of rationality strengthens arguments for the rationality thesis. I shall now consider the merits and demerits of the constrained reading of rationality.

There are at least two serious problems with the notion of constrained rationality. First, although a metric for greater or lesser rationality is required to make sense of the notion of "as rational as possible given certain constraints," it is not at all clear what such a metric would look like. So, for example, supposing that as a result of the human finitary predicament, humans cannot follow the principles of logic appropriate to the selection task; what then is the *most rational* humans could be in the face of their predicament? Should they select only the card showing a vowel to be turned over to test the rule "if a card has a vowel on one side, then it has an even number on the other side" or should they select *both* the card showing a vowel and the card showing an odd number? Once reasoning in accordance with the norms has been ruled out as a possibility, it is hard to see what sorts of considerations can be brought to bear on whether a principle is rational. Of course, there are other considerations for evaluating heuristics besides accordance with the norms, for example, whether the heuristic is quick, whether it will lead to greater reproductive success in a certain environment,[24] and so on. It is not at all clear, however, why any of these considerations are relevant to *rationality*, the matter before us.

Another problem with using the constrained reading of rationality is that it somehow needs to be decided which constraints are to be included in the set of constraints that humans face in virtue of being human. To illustrate this, imagine a possible world in which humans have slightly different brains (from the brains they actually have), and, as a result, can*not* perform a particular reasoning task that in the actual world they can perform. On the constrained reading of rationality, since this brain defect is a feature of the human condition (in the possible world), humans would still be rational even though they diverge from both the norms of reasoning and from actual practice. But if this is right, then humans would be rational on the constrained reading regardless of what their cognitive competence is. The constrained reading of rationality thus turns the rationality thesis into a tautology. If humans do not have a particular normative principle of reasoning in their cognitive competence, then it is open to friends of the rationality thesis who adopt the constrained reading of rationality to say that humans are still rational despite this divergence from the norms because their divergence from the norms is a result of the human condition.

The notion of rationality given the constraints humans face is thus quite problematic. First, it suffers from being unclear, since it lacks a metric for picking out greater rationality, and, second, it makes the rationality thesis tautological, since any principle that does not match the norms can be counted as a constraint in the face of which humans must reason. In light of these problems, in the rest of this essay, I will continue to use the straightforward normative sense of rationality that I introduced above in section III.

[24] The idea that the most rational heuristic is the one that leads to the greatest reproductive success will be discussed in Chapter Five section VIII.

V.

Having said more explicitly what the two opposing theses are and what they are not, I shall now examine the various ways friends of the rationality thesis try to show that the irrationality experiments do *not* establish the irrationality thesis. The three irrationality experiments described above and others like them have a standard form: some supposed norm of human reasoning is shown to be frequently and systematically violated by humans. On the face of it, results of this sort count in favor of the irrationality thesis; in the terms I introduced at the end of section III, the experiments suggest that human cognitive competence does not match the normative standards of reasoning. In general, friends of the rationality thesis respond to these experiments by saying that the results reflect performance errors. There are two ways that they support this response. The first involves focusing on the particulars of the experiment to show that subjects are in some way not really violating the norm in question. The second involves making some general argument to the effect that humans are rational and arguing that it follows from this general argument that any reading of the irrationality experiments that suggests otherwise must be mistaken. I will consider these in turn.

Turning first to the strategy of saying subjects do not really violate the norm, consider the selection task. Typically, subjects in this experiment are seen as interpreting the statement "if there is a vowel on one side of a card, then there is an even number on the other side" as a material conditional with "there is a vowel on one side of a card" as the antecedent and "there is an even number on the other side" as the consequent. This, coupled with the results of the experiments suggest that subjects are seen as making a logical error, a mistake about how to test the truth of a material conditional. To test a conditional, one needs to examine cases in which the antecedent is true (to see if the consequent is false) and cases where the consequent is false (to see if the antecedent is true). Subjects fail to do this.

One might object, however, that this is the wrong way to look at subjects' behavior. Perhaps subjects are in some way misconstruing the instructions of the experiment and thus not classifying the rule they are asked to test as a material conditional; given that 'if-then' statements are notoriously ambiguous in natural language,[25] perhaps some such ambiguity is responsible for the misclassification of the rule. If either of these accounts is the correct way to look at the experiment, then subjects' logical abilities are not necessarily discredited. The same sort of move could be made with the conjunction experiment; perhaps subjects are misinterpreting the sort of judgment that they are being asked to make—instead of *probability* judgments, perhaps they think they are being asked to make *plausibility* judgments (a philosophy major and former civil rights activist becoming a bank teller might seem implausible; the story might seem more plausible, however, if she is also a feminist). Or perhaps subjects are misinterpreting the meaning of one or more of the choices, for example, maybe they think "Linda is a bank teller" means, in tl context of the experiment, that Linda is a bank teller and *not* a feminist. The general strategy is, when faced with a particular irrationality experiment, to argue that subjects are not violating the norm in question because they are misinterpreting some aspect of the task they are asked to perform; if subjects are not doing what experimenters think they are, then subjects are unlikely to be violating the norms they are suspected of violating. I will call this strategy the "misinterpretation strategy."[26] This strategy of explaining the results of the irrationality experiments as misinterpretations on the part of the subjects is perfectly plausible, but it is

[25] Consider, for example, the statement 'If you eat all your broccoli, you will get to have a piece of cake.' If Billy interprets this as a material conditional, he will think that its truth is consistent with him not eating his broccoli but still getting a piece of cake. The statement is meant, in fact, to be read as equivalent to the biconditional statement: 'You will get a piece of cake if and only if you eat all your broccoli.'

[26] Versions of the misinterpretation strategy have been suggested by Mary Henle, "On the Relation Between Logic and Thinking," *Psychological Review* 69 (1962), pp. 376-382; and Henle, "Foreword," in *Human Reasoning*, R. Revlin and R. E. Mayer, ed. (Washington D.C.: Winston, 1978), pp. xiii-xviii. Also see M. D. S. Braine, B. J. Resier and B. Rumain, "Some Empirical Justification for a Theory of Natural Propositional Logic," in *The Psychology of Learning and Motivation*, volume 18, Gordon H. Bower, ed., (Orlando, FL: Academic Press, 1984), pp. 313-371; and L. Jonathan Cohen, "Can Human Irrationality Be Experimentally Demonstrated?"

also open to empirical evaluation—the view involves claims that can be experimentally established. In Chapter Four, I will consider the misinterpretation strategy as well as the evidence for and against it.

A more theoretical argument that subjects in the irrationality experiments should not be interpreted as being irrational involves the principle of charity. W. V. O. Quine recommends that this principle should be followed in translating the utterances of a speaker, namely, that we should translate her utterances in a way that interprets them as reasonable and consistent rather than absurd or contradictory.[27] Others have suggested that this same principle be applied to the interpretation of people's mental states, cognitive heuristics, etc.[28]—if we are to interpret a person's beliefs and mental mechanisms, we must assume they are rational, most of their beliefs are true and most of their heuristics are good detectors of truth. If some strong version of the principle of charity is right for cognitive states and heuristics, then cognitive scientists are mistaken in interpreting subjects in the irrationality experiments as being irrational; if the principle of charity is adopted, then the irrationality thesis is undermined. In Chapter Four, in addition to discussing the misinterpretation strategy, I will examine the principle of charity and how it might be used to combat the irrationality thesis by forcing an alternative interpretation of the irrationality experiments.

Another argument (actually, a *set* of arguments of similar structure) that humans are rational and, therefore, that the results of the irrationality experiments should not be interpreted as demonstrating that humans are irrational takes the analogy between linguistic and cognitive competence and goes one step further. In the first part of Chapter Two, I discuss the attempt to apply the competence-performance distinction to reasoning. In the

[27] W. V. O. Quine, *Word and Object* (Cambridge: MIT Press, 1960); and "Ontological Relativity," in *Ontological Relativity and Other Essays* (NY: Columbia University Press, 1969), p. 46.

[28] See the works of Daniel Dennett, especially *The Intentional Stance* (Cambridge: MIT Press, 1987); the works of Donald Davidson, especially *Inquires into Truth and Interpretation* (Oxford: Oxford University Press, 1984); Elliott Sober, "Psychologism," *Journal of Social Behavior* 8 (1978), pp. 165-191; and Cohen, "Can Human Irrationality Be Experimentally Demonstrated?"

second part of Chapter Two and in Chapter Three, I discuss three arguments that the errors subjects make in the irrationality experiments must be performance errors, not competence errors, that is, not errors that reflect any deep problems in their ability to reason. As such, even the repeated occurrence of such errors is not indicative of irrationality. Seen in this light, the irrationality experiments are useful in that they identify performance errors in reasoning; however, the errors subjects make are not due to any underlying irrationality. Subjects *are* rational—they have a cognitive competence (analogous to linguistic competence) that matches the correct normative standards of reasoning—and the errors they make can be explained in one of several ways, but not in ways that undermine their rationality.

As an example of this, consider again the concrete version of selection task. In the version of the selection task involving the deck of cards with the name of a city on one side of each card and a mode of transportation on the other, subjects generally performed in accordance with the rules of logic. Someone who wants to apply the competence-performance distinction to reasoning could say that the abstractness of the standard version of the selection task is the cause of the performance errors or that the realism of the concrete version facilitates the triggering of the right principle in cognitive competence.[29] According to this argument, in more concrete cases, subjects' cognitive competence is not blocked by any performance static and the rules of reasoning that constitute cognitive competence are thereby unproblematically applied.[30]

There are two claims being made through the invocation of the competence-performance distinction: the first is that the distinction can be deployed in the realm of reasoning in a way that parallels its role in linguistics (this claim is discussed in Chapter

[29] This has been called the facilitation effect. For discussion, see "Facilitation of Reasoning by Realism," and "Pragmatic Reasoning Schemas."

[30] But, as mentioned above, Cosmides, "The Logic of Selection," is among those who offer experiments and analyses that discredit the view that the *concreteness* of the task is relevant to subject's performance. Perhaps a more sophisticated version of the competence-performance distinction could, however, be worked out based on her results. See also *Inference and Understanding*, Chapter Six.

Two); the second is that, assuming the distinction can be deployed in this way, human cognitive competence matches the norms of reasoning and thus humans are properly thought to be rational. The two claims are connected but distinct—the competence-performance distinction might apply to reasoning and yet human cognitive competence could still diverge from the norms of reasoning. Why should we think that the conjunction rule in probability theory is part of our cognitive competence, particularly when Kahneman and Tversky's conjunction experiment seems to show just the opposite?

The first argument that the norms of reasoning are part of our cognitive competence comes from Elliott Sober. He argues that considerations of simplicity favor seeing human cognitive competence as matching the norms of reasoning and explaining the results of the irrationality experiments as performance errors rather than seeing heuristics that diverge from the norms as part of human cognitive competence.[31] The second argument comes from John Macnamara who argues that since our norms of reasoning are based on our considered intuitions about what constitutes good reasoning, and since these considered intuitions come from human cognitive competence, human cognitive competence and the norms of reasoning must match. Macnamara's conclusion is that human cognitive competence consists of only the norms of reasoning and that divergences from these norms should be thought of as performance errors.[32] Third, L. Jonathan Cohen argues, somewhat similarly, that the method for discovering our cognitive competence and the method for justifying the norms of reasoning are identical, and that therefore our cognitive competence cannot fail to match our norms. According to Cohen, the proper method of justifying norms of reasoning is the process of reflective equilibrium. On Cohen's construal of this process, a rule of reasoning is justified if it is the result of a balancing of our accepted reasoning practice and our intuitions about what the norms of reasoning are. But this process, Cohen claims, is isomorphic to the process of discovering what human

[31] "Psychologism," pp. 177-180.

[32] *Border Dispute*, Chapter Two, and pp. 180-185.

cognitive competence is. If the process of discovering what rules of reasoning humans actually follow (performance errors aside) is the same as the process of determining what the norms of reasoning are, then the results of the two processes must be the same and our cognitive competence must match the norms of reasoning. I will consider the arguments of Sober and Macnamara in Chapter Two and turn to Cohen's arguments in Chapter Three.

Another way to undermine the claim that subjects in the irrationality experiments are violating norms and are thereby irrational is by appeal to evolution. The idea is that evolution, through natural selection, produces organisms with mechanisms that select true beliefs (what I call truth-tropic mechanisms or heuristics) and that organisms with such mechanisms would be rational ones. Humans, being the result of natural selection, have truth-tropic mental mechanisms; this provides a *prima facie* reason for not interpreting the results of the irrationality experiments as showing that humans are irrational.[33] This argument, like the argument based on the principle of charity, does not (unlike particular versions of the misinterpretation strategy) specify what the correct interpretation of the irrationality experiments is; rather, the evolutionary argument is supposed to show that the irrationality experiments can*not* establish the truth of the irrationality thesis because the evolutionary history of our cognitive heuristics provides a strong reason for interpreting these heuristics as rational.

There are two possible modifications to the evolutionary argument that I also consider. The first modification is to attempt to connect evolution and rationality through reproductive fitness rather than through truth. This evolutionary argument begins with the claim that

[33] This argument seems to be accepted at least implicitly by Daniel Dennett, "Making Sense of Ourselves," in *Intentional Stance* (Cambridge: MIT Press, 1987); Jerry Fodor, "Three Cheers for Propositional Attitudes," in *Representations* (Cambridge: MIT Press, 1981); Elliott Sober, "Evolution of Rationality," *Sythnese* 46 (1981); W.V.O. Quine, "Natural Kinds," in *Ontological Relativity and Other Essays* (NY: Columbia University Press, 1969); Karl Popper, "Evolutionary Epistemology," in *Evolutionary Theory: Paths into the Future*, J.W. Pollard, ed. (London: Wiley and Sons, 1984); Alvin Goldman, *Epistemology and Cognition* (Cambridge: MIT Press, 1986); William Lycan, "Epistemic Value," in *Judgement and Justification* (Cambridge: Cambridge University Press, 1988); Ruth Millikan, "Naturalist Reflections on Knowledge," *Pacific Philosophical Quarterly* 65 (1984), pp. 315-334; and David Papineau, *Reality and Representation* (Oxford: Basil Blackwell, 1987).

evolution involves selection for reproductive success and the claim that having heuristics and mental mechanisms that lead to reproductive success is all it takes to count as rational. Putting these two claims together, we get the argument that since humans have evolved, they will have mechanisms and heuristics that lead to reproductive success and hence they will be rational.

The second modification draws from an approach to theory of knowledge known as evolutionary epistemology.[34] The two versions of the evolutionary argument considered thus far are concerned with innate heuristics and mental mechanisms. Perhaps, as Stich has recently argued,[35] the heuristics that guide reasoning are *not* innate. If that is the case, then biological evolution and natural selection cannot be the central driving force behind the development of our mental mechanisms. This is where evolutionary epistemology is supposed to come in. On one version of this view, the development of human knowledge is governed by a trial-and-error process analogous to biological natural selection. If the heuristics that govern human reasoning come from a process analogous to biological natural selection (what I call epistemic natural selection), then the evolutionary argument for rationality might be successful. This argument would proceed as follows: epistemic natural selection would select heuristics that produce true beliefs; having such truth-tropic heuristics makes an organism rational; since humans acquire their heuristics through epistemic natural selection (according to evolutionary epistemology), humans are, therefore, rational. I will consider the various versions of the evolutionary argument in Chapter Five.

[34] See, for example, Donald Campbell, "Evolutionary Epistemology," in *The Philosophy of Karl Popper*, volume 1, Paul Arthur Schilpp, ed. (LaSalle, IL: Open Court, 1974), pp. 413-463; and Michael Bradie, "Assessing Evolutionary Epistemology," *Biology and Philosophy* 1 (1986), pp. 401-459.

[35] Stich, *Fragmentation*, pp. 71-74.

VI.

The arguments for the rationality thesis that I discussed above—with the exception of the argument based on the misinterpretation strategy (because this argument is supposed to be based strictly on empirical evidence from cognitive science)—suggest (at least implicitly) that there are limits to the claims cognitive science can make. This is not so surprising since all scientific enquiry is limited in some way. If, however, any of these arguments for the rationality thesis is right, then cognitive science is limited in a particular and, at least to some cognitive scientists, surprising way, namely, cognitive science is limited in such a way that it cannot demonstrate that humans are irrational.

Philosophers quite commonly (and others not quite so commonly) make claims about the limits of science and of particular scientific disciplines. Take, for example, Hume's problem of induction that says one is not justified in reasoning from past occurrences of an event-type to future ones. This implies that no scientific claims about the future are justified (in the strongest sense). So, if the problem of induction is right, science is limited to making knowledge claims about just the past and the present; it cannot be justified in making claims about the future. Or, to take another example, some sociobiologists have tried to make an assortment of ethical and meta-ethical claims based on biological facts, for example, that incest is immoral or that ethics ought to be "biologized."[36] Philosophers have responded that such ethical and meta-ethical claims are not in the purview of biology, that is, biology is limited in such a way that it cannot make ethical and meta-ethical claims. In making such a claim, these philosophers need not commit themselves to an intellectual division of labor between people; that is, they need not claim, for example, that only people with Ph.D.'s in philosophy get to say what is moral and immoral. Rather, all they need to claim is that there are different type of facts in the world—for example, empirical and

[36] See, for example, E.O. Wilson, *Sociobiology: The New Synthesis* (Cambridge: Harvard University Press, 1975).

conceptual ethical and non-ethical, scientific and non-scientific—and that, in the case of biology and ethics, biological facts do not establish ethical claims.

The upshot of the arguments for the rationality thesis that I will be considering in the chapters that follow (with the exception of the one based on the misinterpretation strategy) is quite similar to this. These arguments entail that whether humans are rational is not the sort of empirical question that the cognitive scientists who have performed and analyzed the irrationality experiments think it is. These cognitive scientists claim that their experiments show humans are irrational. If any of the arguments for the rationality thesis (including the one based on the misinterpretation strategy) are right, then the irrationality experiments, at best, isolate an assortment of performance errors that humans are apt to make. If, however, the arguments for the rationality thesis fail, then there is at least *prima facie* reason for thinking that cognitive science in general and the irrationality experiments in particular are relevant to whether humans are rational. Questions about the limits of cognitive science will be in the background throughout the next four chapters and will be in the foreground in Chapter Six.

Chapter Two: Cogitations on Cognitive Competence

I.

In Chapter One, I described the rationality thesis as the view that human cognitive competence matches the normative standards of reasoning. This way of characterizing the rationality thesis draws the idea of a competence from linguistic theory. In this chapter, I will discuss the notion of linguistic competence and whether an analogous notion of competence can be applied in the realm of reasoning. I will also examine three arguments that human cognitive competence matches the correct rules of reasoning (in accordance with the rationality thesis).

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." This notion of a competence is at one level an intuitive notion. We might say that Albert has a competence for playing the piano if he knows how to play the piano well or if, under the right circumstances, he plays the piano well. Perhaps when certain conditions hold, (for example, if he has a broken arm or is under the influence of LSD), we would expect that he would be unable to play at his usual high standards. We would allow for this and even for Albert's just having an occasional bad day, yet still think that he has a competence for piano playing. The idea is that sometimes Albert's actual piano-playing behavior is not up to the standards that characterize his capacity for playing the piano. It is roughly this intuitive sense of competence that is formalized in linguistic theory. Those who wish to use the competence-performance distinction in the realm of reasoning to develop a notion of cognitive competence claim to apply the distinction from linguistics to reasoning unaltered. To see if they do so successfully, I first examine how the distinction is used in linguistics.

II.

The central project of linguistics, as characterized by Noam Chomsky, is to develop an account of the linguistic knowledge of humans, that is, a theory of our knowledge of language.[1] Linguists do not, however, have direct access to this knowledge; our underlying linguistic competence is accessible primarily through actual linguistic behavior such as speech and comprehension. We often, however, make linguistic mistakes—we utter ungrammatical sentences and classify as ungrammatical sentences that are perfectly grammatical. Sometimes these mistakes are due to a lack of attention on the part of the speaker or listener because of an inadequate amount of sleep, excessive alcohol consumption or excitedness (call these situational factors). Other times, these mistakes are due to basic facts about the human condition such as constraints on processing time and memory (call these psychological factors). For example, there are grammatical sentences that would take centuries to utter that no human could judge to be grammatical because of their finite life spans. Mistakes due to either of these types of factors are performance errors; they do not reflect an underlying lack of understanding of the language by the speaker or hearer. Looked at another way, since what linguists want to describe is a specific underlying faculty—the faculty for language use and understanding—that interacts with other faculties, capacities, mechanisms, etc., they necessarily need to idealize, to abstract away from these other abilities, in order to "hone in" on linguistic competence. As Chomsky says:

> Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and

[1] See, for example, Noam Chomsky, *Aspects of the Theory of Syntax* (Cambridge: MIT Press, 1965); *Reflections on Language* (New York: Random House, 1975); *Rules and Representations* (New York: Columbia University Press, 1980); *Language and Problems of Knowledge* (Cambridge: MIT Press, 1980); and *Knowledge of Language* (New York: Praeger, 1986).

> interest, and errors (random or characteristic) in applying his knowledge of language in actual performance.[2]

It may seem strange to abstract from actual behavior this much, but this way of looking at competence is not that different from our commonsense notion of competence. In terms of the Albert example, we might say that Albert is highly competent at playing the piano even if, in fact, he never again plays the piano (say because there are no pianos around or because his hands get cut off) so long as he has the relevant piano-playing knowledge and capacity.

There are various ways of talking about linguistic competence. First, one might connect competence to the operation of a "mental organ"—an underlying mechanism that embodies linguistic knowledge and is responsible for language capacities. Chomsky introduces the general notion of a mental organ in the following passage:

> We may usefully think of the language faculty, the number faculty, and others as "mental organs," analogous to the heart, or the visual system or the system of motor coordination and planning.[3]

Chomsky's analogy between linguistic competence and, say, the visual system is illustrative of several claims. First, the basic structure of both are genetically determined. Second, both will develop in different ways according to environmental inputs—for example, if a child is not exposed to a certain amount of light in her formative years, she will not develop normal visual capacities; similarly, if a child is not exposed to linguistic input, she will develop impoverished language capacities. Third, the ontogenetic development of both will happen without conscious attention to the process—a child does not have to consciously learn to develop eyes; a child does not have to consciously learn to develop linguistic abilities.

Fodor, in *The Modularity of Mind*, points out an important difference between mental organs and anatomical organs that Chomsky glosses over. He writes, "When Chomsky says that there is an innately specified 'language organ,' what he means is primarily that

[2] *Aspects of the Theory of Syntax*, p. 3.
[3] Noam Chomsky, "Rules and Representations," *Behavioral and Brain Sciences* 3 (1980), p. 3.

there are truths (about the structure of possible first languages) that human beings innately grasp."[4] But there are no such truths (not even any beliefs) that humans innately grasp in virtue of having a heart. In other words, mental organs involve *propositional attitudes*, specifically, in the case of the language organ, beliefs about the structure of languages, while organs of the non-mental sort, in contrast, do not involve any innately grasped truths or even any propositional attitudes. This is not an argument against the mental organ picture of linguistic competence, but a warning against potential misunderstandings of it.

On the account of competence that involves the language organ, linguistic competence is the unfettered operation of this organ, that is, the behavior of the language organ if it were attached to the best possible input-output devices, memory, etc.; as an example of this, you might think of extracting the language organ[5] from the brain and implanting into a very powerful computer that would enable it to operate at its maximum capacity.

Note that the analogy between mental and anatomical organs need not be as strong as my discussion so far has suggested for there to be an underlying capacity for language. Presumably, some people have a competence for knitting, an underlying ability to knit that is independent of actual knitting behavior and that can be distinguished from performance errors made while knitting. This could be true without there being a strong analogy between knitting competence and the visual system, that is, without it being the case that knitting competence is genetically determined and ontogenetically unconscious. The point is that there are several versions of the organ way of looking at linguistic competence: on Chomsky's view, the language organ is like the visual system; on the weaker view, it is a mere capacity, like the capacity to knit; and on the third and perhaps most contentious view, linguistic competence is a "module." Fodor defines a module as a domain-specific,

[4] Jerry Fodor, *The Modularity of Mind* (Cambridge: MIT Press, 1983), p. 7.

[5] There is a sense in which this is overly simple—not any language organ will do. Implicit in this sense of competence is that the language organ whose unfettered operation we are interested in is a "normal" language organ.

innately specified, hard-wired and autonomous mental organ and he argues that the language organ is a module.[6]

A second way of talking about linguistic competence is as a person's knowledge of language.[7] Chomsky tries very hard to distinguish between knowledge of language and linguistic ability or capacity. He writes:

> [I]t does not seem to me quite accurate to take "knowledge of English" to be a capacity or an ability, though it enters into the capacity or ability exercised in language use. In principle, one might have the cognitive structure that we call "knowledge of English," fully developed, with no capacity to use the structure [see, for example, Chomsky's imaginary example of Juan who develops aphasia and temporarily loses his capacity for speech and language understanding but not his knowledge of language[8]]; and certain capacities to carry out "intellectual activities" may involve no cognitive structures but merely a network of dispositions and habits, something quite different. Knowledge, understanding or belief is at a level more abstract than capacity. . . . The notions of "capacity" and "family of dispositions" are more closely related to behavior and "language use"; they do not lead us to inquire into the nature of the "ghost in the machine" through the study of cognitive structures and their organization.[9]

This idea of linguistic competence as knowledge of language is potentially confusing in two ways. First, knowledge is typically thought to be a sort of justified belief. But it is far from clear in what sense our linguistic principles are justified. Chomsky in some places tries to avoid this confusion by talking about "cognizing" linguistic principles rather than having knowledge of them.[10] Chomsky has dispensed with this admittedly awkward term but does not, when he talks of knowledge of language, mean to imply knowledge in the strong philosophical sense. Second, although Chomsky wants to distance himself from an

[6] *Ibid.*, p. 37. Like Chomsky's claim that there is a language organ, Fodor's claim that there is a language module is empirical; that there is a language *module* entails that there is a language organ, but the reverse is not true. For a hint that Chomsky disagrees with the view of the language organ as modular, see *Knowledge of Language*, p. 14. See Jay Garfield, ed., *Modularity in Knowledge Representation and Natural-Language Understanding* (Cambridge: MIT Press, 1987) for further discussion of modularity.

[7] Jerry Fodor and Merrill Garrett, "Some Reflection on Competence and Performance," in *Psycholinguistic Papers*, J. Lyons and R. J. Wales, eds., (Edinburgh: Edinburgh University Press, 1966), pp. 135-154, make a distinction similar to that between the two sense of competence I have discussed so far. They characterize the two senses as, on the one hand, the distinction between studying behavior and what underlie it, and, on the other hand, the distinction between linguistic information and the *psychological mechanism* that underlies it.

[8] *Language and Problems of Knowledge*, pp. 10-11.

[9] *Reflections*, p. 23.

[10] *Ibid., passim.*

account of linguistic competence as a disposition towards certain linguistic behaviors, knowledge, particularly of the tacit sort, is hard to understand in any way *except* as a disposition or capacity to cause certain behaviors. Chomsky can agree to this point—human knowledge of language may currently be accessible only through linguistic behavior—while still accepting that knowledge of language is more than just linguistic behavior. For example, perhaps neuroscience will some day be able to access our knowledge of language independent of our linguistic behavior; this would show that knowledge of language is more than an ability or capacity even though it is this knowledge that constitutes the linguistic abilities that we have.

The third way of thinking about competence is as an idealization of behavior. On this sense of the term, competence would be behavior under ideal conditions. This characterization is, as put, ambiguous; it conflates two different ways idealization could take place. On one reading, ideal conditions would be those conditions under which the language mechanism functions properly. On this reading, however, competence as an idealization of behavior just amounts to competence as the operation of the language organ. On the other reading, the idealization would be guided by simplicity considerations, the desire for a unified theory, and the like. So, on this view, linguistic competence is linguistic behavior when the person has had enough sleep, is sober, and so on. This is different than competence as the operation of the language organ. To see this, consider the auditory system and what might be thought of as auditory competence. Suppose that the auditory system was constructed in such a way that there was a "deaf spot," a specific sound frequency, within the range that humans can *actually* hear, that (counterfactually) humans cannot hear. On the competence as the "operation of a mechanism" sense of the term, human auditory competence would, in the imagined case, be characterized as having a deaf spot since, even under ideal conditions, humans cannot hear sounds of that particular frequency. In contrast, on the competence as the idealization of behavior, the deaf spot would be seen as the result of performance errors (or the like) and would be abstracted from because of simplicity and unity considerations; human auditory competence would thus be characterized as continuous, that is, without any deaf spots. The same sort of difference would hold between both the behavior idealization and the operation of the underlying mechanism view of linguistic competence.

Given the "deaf spot" example, the behavior idealization view of competence may seem like it is not a view of competence after all. The linguist sets out to develop an account of the actual human capacity for language; the behavior idealization view, because it places a premium on simplicity, and the like, seems like it will, at best, issue in an idealized approximation to competence rather than actual competence. This is the essence of the difference between the behavior idealization view of competence, on the one hand, and the knowledge and mechanism operation views of competence on the other. Friends of the behavior idealization view might, however, argue that behavior idealization is the best hope we have for developing a theory of competence given our epistemological situation. Since linguists cannot actually access the language mechanism or our knowledge of language, they must rely on what is accessible, namely linguistic behavior. This perhaps confuses the metaphysical question of what our competence *is* with the epistemological question of how what we can *know* about our competence, but I need not pursue this line of argument any further.

So, to summarize, there are three general ways of looking at linguistic competence: as the knowledge of language, as the idealization of linguistic behavior, and as the behavior of the underlying mechanisms responsible for language. The mechanism view of competence itself comes in three flavors: the mechanism can be a module, an organ or a mere capacity. Having surveyed these various senses of competence, for most of what follows, I want to ignore these differences and talk generally about competence as a mental capacity or cognitive system for doing something under ideal circumstances; unless I indicate otherwise, I mean to abstract away from the differences among the various senses of competence. This has the virtue of not linking my arguments to the truth of some particular linguistic theory.

Consider now an objection to the notion of linguistic competence. Since performance errors do occur, linguists cannot read their theory directly off of linguistic behavior. Rather, while they begin their inquiry with a study of actual performance, they must abstract from it to develop a theory of competence. Some actual utterances will be attributed to psychological factors (for example, limits on short-term memory) and situational factors (for example, drunkenness) and not counted as data to be explained by a theory of competence. If, for example, I utter the ungrammatical sentence "I are sitting at my desk," a linguist need not alter her general competence theory or her competence theory of English; rather, she would explain my uttering an ungrammatical sentence as a performance error, for example, as due to my not having had enough sleep.

This use of the competence-performance distinction may, however, seem suspect. A similar sort of distinction could be employed in such a way that would have the effect of making a theory unfalsifiable. Consider, for example, how an astrologer might use an analog of the competence-performance distinction to block evidence from counting against astrological theory. Suppose, for example, that an astrologer claims that the position of the planet Venus indicates that Scorpios will experience romantic trauma on November eleventh. A skeptic of astrology might offer testimony from dozens of Scorpios who experienced no such trauma on the indicated date. The astrologer might respond that this prediction failed due to some "performance" error, such as interference in the orbit of Venus caused by a comet, and thus that the skeptic's evidence does not constitute an objection to astrological theory; astrology would thus remain untainted by such supposed counterevidence. Something should, however, seem suspicious; the astrologer's original prediction made no mention of comets. Though the details of the example are obviously contrived, the point is that some astrologers behave as if they believe nothing should be taken as a falsification of astrological theory.

The general issue here is a standard one in philosophy of science. Karl Popper notes that

> . . . it is always possible to find some way of evading falsification [of a theory], for example by introducing *ad hoc* an auxiliary hypothesis or by changing *ad hoc* a definition. It is even possible without logical inconsistency to adopt the position of simply refusing to acknowledge any falsifying experience whatsoever.[11]

These "immunization strategies"[12] allow any empirical evidence against a theory to be discounted. In the case of astrology, nothing would count as evidence against astrological theory in the eyes of many "practicing" contemporary astrologers. A general problem for the use of the competence-performance distinction is to make sure that the distinction does not act as an immunization strategy.

Popper characterizes an immunization strategy as an unjustified modification (in the sense that there is no justification for the modification other than that it would prevent the theory from being falsified by the data) of a theory in such a way that some particular data will not falsify it. In a sense, this characterization just pushes the question back—what distinguishes such an *ad hoc* adjustment to a theory from a motivated, reasonable one? Answering this question is beyond the scope of this essay so instead of answering this question, I will implicitly appeal to the intuitive difference between an *ad hoc* and an appropriately motivated modification of a theory.

There are three reasons why using the competence-performance distinction is a legitimate strategy. First, the performance errors that linguists appeal to have plausible explanations; they fit a pattern and have a structure. That linguistic competence is not exhibited in atypical situations is unsurprising—chemical changes in the brain due to situational factors such as alcohol in the blood stream should affect linguistic behavior in virtue of the fact that linguistic utterances are initiated in the brain. That linguistic competence is also not exhibited due to psychological factors like memory limitations is

[11] Karl Popper, *The Logic of Scientific Discovery* (New York: Basic Book, 1959), p. 42.

[12] Karl Popper, *Objective Knowledge: An Evolutionary Approach* (Oxford: Oxford University Press, 1972), p. 30.

similarly predictable. Linguistics, like other sciences, should be expected to make use of *ceteris paribus* clauses, perhaps ineliminable ones.[13] (This suggests a natural way to look at performance errors: they are what happens when *ceteris* is not *paribus*.) Performance errors in linguistics thus fit into a general picture of human psychology; the application of the competence-performance distinction is not unmotivated like an immunization strategy.

A further reason for thinking that using the competence-performance distinction is a legitimate strategy is that linguists do not typically appeal to this distinction when the data does not fit with their favorite theory. Chomsky, for example, has made major changes to his preferred theory of linguistic competence over the years, changes that he could well have avoided making if he had used the competence-performance distinction as an immunization strategy. That Chomsky chose instead to modify in great detail his theory of competence so that it would fit the data suggests that the competence-performance distinction is not a linguist's "vaccine."

Finally, the competence-performance distinction is important to linguistics because even if people never made any performance errors, a theory of linguistics would need to go beyond actual linguistic behavior to develop a full theory of human linguistic knowledge. There are an infinite number of sentences that will never be uttered. A complete linguistic theory needs to include these possible utterances as part of our linguistic knowledge. A theory of performance based only on actual utterances would be unable to do so. A theory of competence is required for a complete linguistic theory. The competence-performance distinction is, thus, quite important for the development of linguistic theory. As such, it does not seem to be an immunization strategy. So, some version of the competence-performance distinction seems a legitimate and perhaps necessary tool for linguists to use.

[13] On the ineliminability of *ceteris paribus* clauses, see, for example, Nancy Cartwright, *How the Laws of Physics Lie* (Oxford: Oxford University Press, 1983).

III.

Given the intimate connection between language and cognition as well as the seeming usefulness of the competence-performance distinction for the study of language, exploring the applicability of this distinction to the study of cognition in general seems a natural and promising course of study. The general idea of applying the competence-performance distinction to reasoning is to distinguish between inferences made in accordance with a person's underlying cognitive competence and those resulting from performance errors. Actual human reasoning behavior would be explained as the operation of an underlying cognitive competence and performance errors due to situational or psychological factors. Cognitive competence would be, like linguistic competence, a capacity, specifically, a capacity for reasoning. Cognitive competence can be thought of as m ıde up of heuristics, general rules that guide cognition, for example:

> **the conjunction elimination heuristic**: if you believe the statement "A and B," then believe the statement "A."

Saying the conjunction elimination heuristic characterizes human cognitive competence amounts to saying that, unless certain psychological or situational interferences occur, humans reason according to this rule. This is parallel to the claim that linguistic competence can be characterized by abstract linguistic heuristics such as

> never move a 'wh'-word across both a noun phrase boundary and a sentence boundary at the same time.[14]

Saying this rule is part of human linguistic competence is compatible with the fact that humans sometimes (namely, when there are certain psychological or situational interferences) utter sentence like "Who does John believe the claim that Mary saw?" that

[14] This example, like most of the linguistic examples I offer, is, at best, outdated; I use it for simplicity's sake. I am thus following in the great tradition of offering anachronistic philosophical examples, the best known of which involves the claim that C-fibers are the neurological basis for pain. I trust (a) that some one who knows the contemporary linguistic vocabulary of parameters, theta roles, etc., (that is, some one other than me) could provide examples that are in vogue and (b) that filling in such examples would not alter the substance of my argument.

violutes the rules of wh-word movement.[15] Since the use of the notion of cognitive competence parallels (perhaps even partially encompasses) that of linguistic competence, and since the notion of linguistic competence seems a legitimate one, the notion of cognitive competence seems a legitimate one as well.

Is this all that needs to be said to show we can deploy a notion of competence in the realm of reasoning that is analogous to linguistic competence? In a sense, yes, and, in a sense, no, depending on how strong an analogy is supposed to hold between cognitive and linguistic competence. In linguistics, there is both a descriptive and a normative component to competence; the *descriptive* component is the claim that linguistic behavior can be characterized by invoking the competence-performance distinction, while the *normative* component is the claim that linguistic competence is properly seen as sanctioning only correct linguistic behavior were it not for various interferences. When a linguist offers an account of linguistic competence, she is giving a description of the capacity for language *and* (thereby) she is articulating what the norms of human natural language are. This is different from, say, the neuroscientist who is giving an account of the human auditory system. Such an account describes the workings of the auditory system, but it does not suggest that this is the *right* auditory system; if it did, such an account would be, for example, committed to the view that the right auditory system is one that is not sensitive to high-pitched sounds. The study of linguistic competence is thus normative in the way that the study of the human auditory system is not.

Matters are actually trickier than this, because there is a normative part to the descriptive component of linguistic competence, and I want to distinguish this normative part of the descriptive component from the *primary* normative component of linguistic competence. Any account of a competence involves some idealization and, as such, is normative—

[15] Roughly, the question-sentence "Who does John believe the claim that Mary saw?" is an ungrammatical sentence resulting from transforming the sentence "John believes the claim that Mary saw X," using a so-called wh-word movement rule in an "illegal" fashion by moving 'who' from both the phrase "that Mary saw X" and the sentence "John believes the claim . . . ."

distinguishing between (actual) behavior and (idealized) competence involves, at least implicitly, making normative claims. This is the normative part of the descriptive component, *not* the (primary) normative component that I want to focus on. The primary normative component of linguistic competence is that, by giving an account of linguistic competence, one is (at least implicitly) saying what counts as correct human linguistic behavior. The normative component of linguistic competence that I want to focus on is *not* a part of every characterization of a competence. In general, it is possible for the competence-performance distinction to be applicable to some realm of behavior without having to explain the behavior as the result of a competence that sanctions only correct behaviors coupled with performance errors that cause deviations from this competence. In other words, it is possible for the descriptive component (with its normative part) of the competence-performance distinction to be applicable to some realm in which the normative component is not. A *weak* analogy between linguistic competence and cognitive competence would hold if just the descriptive component of linguistic competence applies to cognitive competence; a *strong* analogy would hold if *both* the descriptive and the normative components of linguistic competence apply to cognitive competence.[16]

Consider, for example, how the competence-performance distinction might apply to ethics. Suppose, as seems true, that humans have an ability to make ethical judgments. No doubt, this ability may be interfered with in various ways—for example, alcohol consumption—so it mig^ht be useful to talk of humans as having an ethical competence (even though this competence may well not be a mental module, a mental organ, and it may not be predominantly innate[17]). Further, we can imagine that this ethical competence might

[16] This talk of the (primary) normative component to linguistics may cause some linguists to cringe as they typically shun the notion of any normative component to their discipline. When they do this, however, they mean that linguistics should avoid the temptation to be prescriptive, for example, to say that English speakers should not use the contraction 'ain't.' But even when linguistics avoids being prescriptive in this sense, it is still normative in the sense with which I am concerned—that is, the sense in which linguistics is normative but the study of the human auditory system is not.

[17] Although, for arguments that there is an underlying ethical capacity that is innate, see, for example, Edward O. Wilson, *Sociobiology: The New Synthesis* (Cambridge: Harvard University Press, 1975); Wilson and Charles Lumsden, *Genes, Minds and Culture* (Cambridge, Harvard University Press, 1981);

match the correct ethical theory or that it might diverge from it, namely, we can imagine some one having an underlying ethical competence that, performance errors aside, sanctions the morally right actions as well as some one whose ethical competence often sanctions immoral actions. If our ethical competence matched the correct ethical theory, then both the descriptive and the normative component of the analogy with linguistic competence would hold and, therefore, so would the strong analogy between linguistic competence and ethical competence. But if this ethical competence diverges from the correct ethical theory, only the descriptive component, and thus only the weak analogy, would apply. In general, if the competence-performance distinction is applied to a realm and if giving an account of competence in that realm involves saying that competence matches the norms, then the strong analogy holds with linguistics; otherwise, if the (primary) normative component is not involved, only a weak analogy holds.

Special care needs to be taken in the realm of reasoning not to conflate the descriptive and normative components of competence. To give a descriptive account of cognitive competence would be to offer an empirically discoverable set of heuristics that characterize the underlying human capacity to reason. In contrast, to give a normative account of cognitive competence would be to offer a set of heuristics that characterize an ideal reasoner. While it might be the case that these accounts would, as some argue, be the same (as their analogs are in linguistics), it begs the question against the irrationality thesis to *assume* that cognitive competence provides both a description of what heuristics humans actually use as well as a prescription of what heuristics an ideal reasoner should use (that is, to assume a strong analogy holds between linguistic and cognitive competence).[18] Making such an assumption would be tantamount to using the competence-performance distinction as an immunization strategy in reasoning. If it is *assumed* that a strong analogy

Promethean Fire (Cambridge, Harvard University Press, 1983); and Michael Ruse, *Taking Darwin Seriously* (Oxford: Basil Blackwell, 1986).

[18] Jay Garfield, "Review of *A Border Dispute*," *Journal of Symbolic Logic* 53 (March 1988), p. 316, claims that John Macnamara, in *Border Dispute: The Place of Logic in Psychology* (Cambridge: MIT Press, 1986), makes just such an assumption by conflating these two senses of cognitive competence.

holds between linguistic and cognitive competence, then any empirical evidence that humans diverge from the norms of reasoning will *a priori* be explained as performance errors. So, for example, the results of the irrationality experiments would *a priori* count as performance errors and hence as not characteristic of human cognitive competence. As such, the competence-performance distinction would be operating as a paradigmatic immunization strategy.

While it is easy to show a weak analogy holds between linguistic competence and cognitive competence—that is, that the descriptive component of the competence-performance distinction applies to both linguistic and reasoning behavior—it is harder to show a strong analogy holds between them—at least, *prima facie*, there may be a difference between the norms of reasoning and the competence underlying human reasoning behavior. This may seem more plausible if compared to an example from a mundane realm. Consider the application of the competence-performance distinction to the ability to drive a car. The general idea of characterizing driving behavior would be to distinguish between behaviors based on underlying driving competence and those based on various interfering factors. Note that doing so is enough to establish a weak analogy with linguistic competence; making the strong analogy, however, requires good reasons for thinking that the underlying driving competence is equivalent to *ideal* driving behavior. But this requirement seems like it would be quite difficult to meet. This is not to say that there is no hope of working out a strong analogy between linguistic competence and driving competence but just that a strong analogy requires further argument. Although there is more of an intuitive pull to the idea that underlying cognitive competence is equivalent to ideal reasoning behavior than to the parallel idea with respect to driving, the point about driving competence also applies to a strong analogy between linguistic competence and cognitive competence: further arguments are required to make the claim; intuitive pull is not enough here. In the following sections and in the next chapter, I discuss just such arguments.

IV.

According to the competence-performance model of language, human linguistic competence involves complete knowledge of grammaticality. Errors occur when other systems with which the language faculty interacts interfere with its operation or when conditions are not ideal for linguistic competence to be displayed. Elliott Sober has suggested that, by analogy (by which he means what I call a *strong* analogy), cognitive competence should be seen as perfectly rational and the results of the irrationality experiments should be explained as due to interference from other systems with which cognitive competence interacts. Sober thus attempts to provide just the sort of argument needed to demonstrate that there is a normative component to cognitive competence analogous to the normative component of linguistic competence. Further, according to Sober, it follows from this claim that the results of the irrationality experiments are consistent with the rationality thesis

> . . . in the same way that grammatical slips are consistent with viewing native speakers as having internalized the grammar of their language. In both cases, one posits a mechanism which endows people with perfect rationality or with perfect grammaticality, and then one posits additional various devices which provide interferences with the smooth functioning of the basically 'correct' mechanism. Instances of ungrammatical utterance or irrational belief and behavior are to be explained as the impingements of interferences like lapses of memory, headaches, substantive prejudice, and so on.[19]

Call Sober's model the perfect competence model of cognitive competence. Like the model of linguistic competence as embodying complete knowledge of grammaticality, the perfect competence model of cognitive competence is an empirical theory, in this case, a theory of what principles characterize our reasoning behavior. If this model is right, the strong analogy between linguistic competence and cognitive competence would hold and the irrationality thesis would be false—the behaviors brought out by the irrationality experiments would all be properly construed as performance errors and human cognitive competence would emerge untainted by irrationality. This model does not beg the question

[19] Elliott Sober, "Psychologism," *Journal for the Theory of Social Behavior* 8 (1982), pp. 177-178.

against the irrationality thesis, that is, used in this way, the competence-performance distinction is not obviously acting as an immunization strategy. With this model, Sober is offering both an empirical theory that counts against the irrationality thesis and considerations that support this theory; Sober is offering just the sort of further arguments that, in the last section, I argued are required for thinking that cognitive competence has a normative component. His model has the additional virtue of unifying language with cognition by showing how both fit under the model of perfect competence.

Sober gives two reasons why the perfect competence model should be favored over its competitor, the imperfect competence model, the view that human cognitive competence is properly characterized as sanctioning erroneous inferences (in other words, the view that only the weak analogy holds between linguistic competence and cognitive competence because the former has both a normative and descriptive component while the latter has just a descriptive component).[20] First, he says that the perfect competence model is

> preferable . . . because the 'errors'—the lapses in rationality and the occurrence of deviant [linguistics] utterances—are occasional and unsystematic . . . ; an irrational logic is plausibly posited only if there is fairly widespread and uniform irrationality at the level of behavior.[21]

Second, he says that the perfect competence model should be preferred for reasons of simplicity. He writes:

> Simple irrational rules of inference (like one sanctioning the fallacy of affirming the consequent) may predict more errors than are consistent with human behavior. Perhaps the only irrational rules of inference whose psychological reality is consistent with actual human performance will be extremely complex. If so, simplicity may favor the model of rational mechanism subject to interferences [that is, the perfect competence model].[22]

Because deviations from perfect competence are unsystematic and because simplicity considerations favor the perfect competence model, Sober concludes that this model is to be preferred to the *imperfect* competence model.

[20] See *ibid.*, p. 178 for Sober's description of this alternate model.
[21] *Ibid.*
[22] *Ibid.*, pp. 179-180.

Despite the initial plausibility of this account, both of Sober's reasons for preferring the perfect competence model are mistaken, and there are further reasons why this model should be rejected. First, the irrationality experiments, if they show anything interesting, show that errors of reasoning occur systematically and with regularity.[23] For example, the conjunction fallacy experiments show that humans systematically and regularly ignore the conjunction rule in making probability judgments. Even from Sober's perspective, the regular and systematic occurrence of errors like these suggests that human cognitive competence includes non-rational heuristics. Sober's criterion that an error be systematic in order for it to be included as a heuristic that is part of cognitive competence is a reasonable one—in the case of cognitive competence, however, the errors the irrationality experiments uncover *are* systematic, so they count as evidence against a perfect competence model of cognitive competence.

Sober, though he does not discuss actual cases like the irrationality experiments, does (at least implicitly) address the possibility that irrational inferential practices might occur systematically. He does this by arguing that considerations of simplicity suggest even systematic occurrences of such errors ought to count as performance errors. Once again, Sober draws from linguistics. He discusses what considerations determine whether uniform cognitive lapses should be attributed to the normal operations of an *imperfect* competence model of cognitive competence or to performance errors occurring due to interference with the operations of a *perfect* competence model of cognitive competence. He compares this to Chomsky's discussion of the rule:

$(\text{if})^n$ snow is white (then grass is green)^n; where $n \geq 0$.

There are an infinite number of strings generated by this rule nearly all of which would be judged ungrammatical by English speakers. In fact, only "Snow is white," "If snow is

[23] Some, for example, Gerd Gigerenzer, "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases,'" *European Review of Social Psychology* 2 (1991), pp. 83-115, have recently argued that the irrationality demonstrated by the irrationality experiments is not at all as systematic as Kahneman, Tversky, and company would have us believe. I will consider some of these arguments in Chapter Four.

white then grass is green," and *maybe* "If if snow is white then grass is green then grass is green" would be accepted.[24] Chomsky, however, claims all of the strings generated by this rule should count as part of the linguistic competence of English speakers. That sentences generated by this rule when n is greater than two are *not* accepted by English speakers is, according to Chomsky, due to the psychological factor that humans have limited computational space for sentence processing. Chomsky argues that, for reasons of simplicity, this general rule is preferable to the explanation based on a rule that simply generates the two or three acceptable instances of the more general rule. Roughly, the idea is that a general linguistic rule is simpler than either the same rule with constraints added (for example, the rule (if)^n snow is white (then grass is green)^n; where $0 \leq n \leq 2$) or rules in the form of lists of acceptable linguistic utterances. More specifically, the general rule is simpler than the rule with constraints because the rule with constraints requires that the language organ has a counter while the general rule does not. The general rule is simpler than a list of acceptable sentences because such lists would take up too much memory.

I will not attempt to assess Chomsky's use of simplicity considerations in linguistics since my concern here is whether considerations of simplicity should, as Sober wants, be applied in the case of cognitive competence and whether, so applied, they provide reasons for embracing the perfect competence model with respect to human cognitive competence. I shall argue (contra Sober) that it is not clear whether the perfect or imperfect model of human cognitive competence is simpler. Consider the conjunction fallacy experiment (that is, the Linda experiment) discussed in Chapter One. Recall that subjects rate the probability of Linda being a bank teller *and* a feminist as higher than the probability of her being a bank teller (whether or not she is a feminist). The perfect competence model of cognitive competence would claim that, despite the seeming experimental evidence to the contrary, the conjunction rule (the probability of A and B is less than or equal to the probability of A)

[24] *Ibid.*, p. 179. For Chomsky's discussion, see *Syntactic Structures* (The Hague, The Netherlands: Mouton, 1957), pp. 23-25.

is part of human cognitive competence and then explain the results of the experiment by saying subjects are making performance errors, errors not indicative of their underlying cognitive competence. In contrast, the imperfect competence model of cognitive competence would claim the conjunction rule is not part of human cognitive competence; rather, this model would describe some other heuristic that is supposed to be part of human cognitive competence and that accounts for subjects' divergence from the norm of the conjunction rule. For example, subjects might be following the *plausibility* heuristic, which says the more *plausible* event (of some set of events) should be believed to be the more *probable*. This heuristic clearly diverges from the norm because sometimes the more plausible event is *not* the more probable one, for example, it may be more *plausible* that Linda is a feminist and a bank teller than that she is a bank teller (whether or not she is a feminist) even though it is not more *probable*. If Sober is right that simplicity considerations always favor the pe.fect competence model over the imperfect competence model, then simplicity considerations ought to favor the conjunction rule account of the conjunction experiment over the plausibility heuristic account. But it is not at all clear that the conjunction rule is simpler than the plausibility heuristic. For example, the perfect competence model requires a story about psychological or situational factors causing performance errors that prevent the proper application of the conjunction rule. Is this more complex than the imperfect competence model that only involves a single heuristic and no such exceptions? The answer is not obvious.

This particular example further counts against Sober's conjecture that the only irrational heuristics that fit with human behavior will be "extremely complex."[25] There seems to be no good reason for thinking that cognitive heuristics that sanction behavior that diverges from the norms will be more complex than those that sanction behavior in accord with the norms—at least, Sober does not offer such reasons.[26] This, however, is not a knock-

[25] *Ibid.*, p. 180.

[26] Much of Sober's discussion around these points seems confused. As an example of the perfect competence model, Sober discusses human visual judgments. Humans are subject to various visual

down argument against the perfect competence model of cognitive competence; it only is an argument against Sober's defense of it.

Before continuing, I want to make a further comment on the simplicity argument for the perfect competence model of cognitive competence. Its failure to produce a clear choice between the perfect and imperfect competence model aside, applying simplicity considerations in this case seems unmotivated. What reasons do we have for thinking that the *simplest* account of cognitive competence is the right one? To the extent that we have reason to think that some parts of cognitive competence—like our linguistic competence—are innate, we have reason to think that our cognitive competence is not so simple. This is because evolution rarely produces simple structures; it is a piecemeal process, a "satisficer." This point will be discussed in Chapter Five below. For now, suffice to say that not only does it seem that simplicity will not give us good reason to think a strong analogy holds between linguistic competence and cognitive competence (and hence that the perfect competence model applies to cognitive competence and the rationality thesis is true), but there seems to be no strong reason in the first place to adopt simplicity considerations with respect to cognitive competence.

Despite the initial plausibility of a Sober-like argument for the strong analogy between linguistic competence and cognitive competence, there is an important difference between linguistics and reasoning that leads to a disanalogy between linguistic competence and cognitive competence thereby creating serious difficulties for the perfect competence model of cognitive competence and the related argument for the rationality thesis. Suppose the human brain was constructed in such a way that linguistic competence was altered in a

illusions. The occurrence of these illusions are, according to Sober (and here I concur), explained as follows:

> Various cues represented in the information extracted by retinal fixations are interpreted in a way that is usually highly reliable. But the particular stimulus conditions considered [in cases of visual illusions] are unusual, and reliable procedures lead the perceiver to formulate incorrect perceptual judgments. (*Ibid.*)

I disagree, however, with Sober's claim that such a story is a case of a rational competence model. The heuristics that we use to interpret visual inputs are fallible (that is, they are not *ideal* visual mechanisms) but they usually do a good job. These visual heuristics are thus like the plausibility heuristic and not like the conjunction rule; without an account of the sort of performance errors that might systematically cause such illusions, our susceptibility to visual illusions suggests we have *im*perfect visual competence.

small but non-trivial 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 type A patterns would *be* ungrammatical and type B patterns would *be* grammatical. In other words, there is nothing grammatical about linguistic patterns that are in fact grammatical (for example type A linguistic patterns) or ungrammatical about linguistic patterns that are in fact *un*grammatical (for example type B linguistic patterns) independent of the actual structure of the human brain. The perfect competence model of linguistic competence applies in the imagined case just as well as it does in the actual case.

Consider the same sort of case in reasoning. Suppose, as Sober thinks, the perfect competence model properly characterizes human cognitive competence. Now imagine the brain were constructed differently and certain heuristics humans in fact follow were, as a result of the different brain structure, not followed (call these type A heuristics). Further, imagine certain other heuristics that humans in fact do *not* follow were, as a result of the different brain structures, followed (call these type B heuristics). If this were the situation, someone who (like Sober) thinks that humans are *in fact* rational would have to say that humans, in this counterfactual situation, are not rational. This is because, by embracing the rationality thesis, one is accepting that the principles of reasoning that 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. The disanalogy is this: linguistic norms (principles of grammaticality) are relative to linguistic competence, while norms of reasoning (principles of rationality) are *not* relative to 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.[27] 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 Sober's view that cognitive competence is properly characterized by the perfect competence model and the related argument for the rationality thesis. This model is well-suited to linguistics because the norms of grammaticality are indexed to actual linguistic competence. In reasoning, however, the norms are not indexed to actual cognitive competence. With the breakdown of this analogy, and in the face of the results of the irrationality experiments, Sober provides no reason to think cognitive competence should be characterized as perfect cognitive competence.

The failure of Sober's arguments for the perfect rationality model of cognitive competence, however, shows neither that something is inherently mistaken about the notion of cognitive competence nor that cognitive competence must diverge from the norms of reasoning. The idea that there is an underlying cognitive competence we have access to through the observation of actual behavior remains a reasonable one; in the terms of section III above, the failure of the strong analogy does not entail the failure of the weak one.

V.

Sober attempts to use simplicity considerations to show human cognitive competence must match the normative standards of reasoning. John Macnamara makes another argument for the same conclusion that deploys yet another analogy with linguistics.

[27] Jay Garfield, "Review of *A Border Dispute*," p. 315, underscores this difference between linguistics and reasoning as follows:

> The assumption that speakers are competent with respect to their native language is akin to, say, the assumption that a sample of gas in a cylinder is ideal—we know that there will be deviations from the ideality [sic] we posit, but they will be sufficiently minor so as not to vitiate the explanatory utility of the competence we ascribe. If the data fail to confirm that sufficiently many speakers make use of the grammar we posit as describing linguistic competence, we revise the grammar, just as if it turned out that most gases fail even to approximate the Boyle-Charles law, we would reject the law. The discovery that most reasoners reject *modus ponens*, or more realistically . . . that they reject Bayes's theorem, does not [in contrast] lead us to revise our logic or probability theory.

Macnamara's argument, as stated in *A Border Dispute*, focuses on what he calls "logical competence," humans' underlying logical ability. His main thesis is that logic provides a model of human logical competence in the same way that the correct linguistic theory provides a model of human linguistic competence. His argument extends quite naturally to cognitive competence and the norms of reasoning writ large. In a nutshell, Macnamara's argument is as follows:

> (1) The (correct) rules of logic are based on our considered logical intuitions.
> (2) Our considered logical intuitions are based on our logical competence.
> (3) Therefore, the (correct) rules of logic are based on our logical competence; given this, human logical competence must match the rules of logic.

The more general argument in terms of reasoning is:

> (R1) The norms of reasoning are based on our considered intuitions about what constitutes good reasoning.
> (R2) Our considered intuitions about what constitutes good reasoning are based on our cognitive competence.
> (R3) Therefore, the norms of reasoning are based on our cognitive competence; given this, human cognitive competence must match the norms of reasoning.

This argument is influenced by a particular account of the relationship between linguistic norms, linguistic competence and linguistic intuitions. As Macnamara sees it (and here he claims to basically follow Chomsky[28]), linguistics operates in the following fashion: the linguist begins by looking at actual linguistic performance (which includes people's intuitions about grammaticality) because there is nothing else besides performance to go on.[29] The linguist abstracts and idealizes from linguistic performance and then uses her own and other people's linguistic intuitions to evaluate the abstracted, idealized candidate rules. The rules that result from this process are the (normative) principles of language; but note they also constitute a characterization of human linguistic competence.[30] This picture of how linguistic competence and linguistic norms connect is supposed to parallel the connection between logical competence and the rules of logic; this picture is behind the argument from (1) and (2) to (3) as well as (R1) and (R2) to (R3).

[28] *Border Dispute*, pp. 22-42.
[29] *Ibid.*, p. 25.
[30] *Ibid.*, p. 26.

Consider first the claim that the rules of logic are based on our considered logical intuitions (1). The way we decide whether a candidate inference is sanctioned by the rules of logic is to check our carefully considered intuitions to see ıf the general form of that inference fits our logical intuitions. Macnamara writes:

> If we have it that "it is raining" and if we also have it that "if it is raining, then the ground is wet," then we are safe in concluding that "the ground is wet." If logicians satisfy themselves that all inferences of this form are valid and that there cannot be an invalid one, then they are entitled to give the common form of such arguments in an abstract schema and state that it is a valid rule of inference Their judgment rests ultimately on logical intuitions . . . [31]

This is supposed to work the same way linguistics does; the linguist notices that strings of words fitting a particular pattern are grammatical and then satisfies herself that all such patterns are grammatical.

Why, one might wonder, are our rules of logic based on our *considered* intuitions? Here, again, Macnamara takes a hint from linguistics. Most people would, at first glance, judge that sentences generated by "center embedding" are ungrammatical, for example, most would say that the sentence

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

is ungrammatical. However, 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 who the cat scratched." So, we now have:

> The girl who the cat scratched fled.

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

> The girl who 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

[31] *Ibid.*, p. 32.

> The girl who 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 re-convince myself of its grammaticality.

The same sort of point is supposed to be true with reasoning. In the conjunction experiment, the probability that Linda is a bank teller *and* a feminist might seem greater than the probability that Linda is a bank teller (whether or not she is a feminist), but, on reflection, we are supposed to realize that this cannot be the case. This way of looking at the irrationality experiments gives support for the view that the norms of reasoning are based on our consic ered intuitions about what counts as good reasoning (R1).

The next premise is that our considered intuitions are derived from our logical competence (2). Roughly, the idea is that when we are careful and reflect on our intuitions, we insure there is not interference, that is, that we are not committing any performance errors. Thus, it basically follows from the definition of competence that considered intuitions provide a window into human cognitive competence. In the realm of reasoning, the claim is that our considered intuitions about what constitutes good reasoning are derived from our cognitive competence (R2). (2) and (R2), coupled respectively with (1) and (R1)—the claims that norms of logic and reasoning are based on considered intuitions—are supposed to entail that the norms of logic and reasoning will match, respectively, human logical competence and human cognitive competence.[32]

This argument for the rationality thesis suffers from the same defect as Sober's argument discussed above in section IV. That argument fails because it does not adequately take into consideration a crucial difference between linguistics and reasoning: grammaticality is indexed to human linguistic competence but rationality is not indexed to human cognitive competence. Whatever our considered intuitions about grammaticality

[32] Several discussions with Paul Bloom helped me get clearer on Macnamara's version of this argument for the rationality thesis.

are, they constitute our competence; not so with our considered logical intuitions. Unlike our intuitions about what constitutes a grammatical sentence, our reflective intuitions about what constitutes a rational inference may well diverge from the norms. For example, subjects in the selection task experiments, even when they are told a card showing an odd number might have a vowel on the other side hence violating the rule "if a card has a vowel on one side of it, then it will have an even number on the other," still insist they were right not to select the odd-numbered card to be turned over.[33] This is just a suggestive example of the obvious point that it is quite possible for people to accept irrational rules of reasoning even on reflection. Given this, the account of the relationship between human cognitive competence and the rules of reasoning embodied in premises (1) and (R1) is simply false.

Why, one might wonder, did premises (1) and (R1)—the claims that the rules of logic and the norms of reasoning, respectively, are based on our considered intuitions—seem plausible in the first place? I think it is due to a confusion between the context of *discovery* and the context of *justification*. While it may be true that our considered intuitions about what constitutes good reasoning play some role in our coming to *discover* (that is, believe) certain principles of reasoning, these intuitions do not necessarily play a role in the *justification* of these principles. These premises seem plausible only to the extent that we conflate how we in fact come to believe something and what *justifies* our belief in it.

A defender of Macnamara's argument for the rationality thesis might respond that I am assuming an account of how our norms of reasoning are developed, how they are justified, and so on. Without such an account, a defender of Macnamara's argument might claim, the difference between linguistics and reasoning on which I put so much weight collapses. In other words, the charge against me is that when I say the norms of reasoning are not indexed to human cognitive competence, I am appealing to a picture of the development and justification of the norms of reasoning I have not articulated or defended. To this charge, I

[33] P. C. Wason and P. N. Johnson-Laird, *Psychology of Reasoning* (Cambridge: Harvard University Press, 1972), pp. 196-197. These results are discussed in greater detail in Chapter Four.

must admit that I am temporarily guilty, but with an explanation and a request for a change in venue. As to the explanation, for my purposes, I do not need to articulate a detailed account of the justification of norms; all I need to demonstrate the failure of the argument for the rationality thesis currently under consideration is that it requires that the norms of reasoning are dependent on facts about human reasoning ability. Since this is, at best, *prima facie* implausible, the burden of proof is on friends of this argument to develop an alternative account, an account they have so far failed to provide. As to the change in venue, in Chapter Three, I will examine at length L. J. Cohen's attempt to provide just such an account—one based on the "reflective equilibrium" model of justification—as part of an argument for the rationality thesis. Cohen's argument is similar to the present one in many ways but Cohen's has the virtue of offering an account of the justification of the norms of reasoning, one in which human cognitive competence plays a role. Without such an account, the argument from (R1) and (R2) to (R3) fails. Whether a similar argument works when such an account is provided will be considered in Chapter Three. In that context, I will consider an alternative account as to where our norms of rationality come from.

VI.

Macnamara suggests another argument for the rationality thesis that is connected to the notion of cognitive competence. He argues for the rationality thesis by what might be thought of as an empirical *reductio ad absurdum*. He writes:

> . . . note that the same set of implicators [that is, mental mechanisms for drawing inferences,[34] in my terms, innate cognitive heuristics] must be available to Kahneman and Tversky on the one hand and to their subjects on the other. How, then, could Kahneman and Tversky use such untrustworthy devices to attain such certain results as the mathematics against which they interpret their subjects' responses? . . . The existence of that mathematics and Kahneman and Tversky's access

[34] *Ibid.*, p. 37.

> to it undermines their rejection of the mathematics as the appropriate competence theory.[35]

The idea is that interpreting the results of the irrationality experiments as establishing the irrationality thesis in a sense proves too much. If, for example, the conjunction fallacy experiment suggests humans have a poor grasp of probability theory, how can the experimenters performing the conjunction experiment be confident about their conclusions that are, of course, based on probability judgments? This *reductio* argument can be laid out as follows:

> (1) For the irrationality experiments to prove any conclusion, the experimenters who perform them must be rational.
> (2) The irrationality experiments show that humans are irrational.
> (3) Since the experimenters are human, they too are irrational.
> (4) Therefore, the irrationality experiments cannot prove any conclusion.
> (5) Therefore, the irrationality experiments do not show that humans are irrational.

This argument, I think, fails.

The irrationality thesis does not make the claim that *every* inference a human makes is irrational; it only makes the claims that *some* of the heuristics humans follow are irrational, for example, humans fail to follow the conjunction rule in certain situations in which they should follow it. For the *reductio* argument against the irrationality thesis to work, it must be the case that the particular rule experimenters claim their subjects fail to follow is invoked in a relevantly similar context by experimenters when they are analyzing the experiment. The conjunction experiment does not show humans *always* ignore the conjunction rule; it may, however, show that humans ignore the conjunction rule in certain sorts of situations. But we have no particular reason to believe that, for example, Tversky and Kahneman must apply the conjunction rule in the same context in which they show humans typically fail to apply it; this is what needs to be shown to undermine the irrationality thesis. In terms of the argument for (5), both (1) and (2) need to be rewritten: (1) should say the experimenter must be rational in the context of the various calculations, inferences, etc., involved in collecting and analyzing the data while (2) should say the

[35] *Ibid.*, p. 184.

irrationality experiments show humans are irrational in the sense that there are various contexts in which they fail to follow the norms of reasoning. With (1) and (2) appropriately revised, the *reductio* will not go through.

This point can be made in a somewhat different way. Consider the claim that for the irrationality experiments to establish something, the experimenters who perform them must be rational (1). In this premise, "rational" can either be interpreted narrowly or broadly. On the narrow reading, the experimenters are rational if the inferences they actually make in the process of reaching their experimental conclusion are rational—that is, if, in evaluating the experimental data, they follow the conjunction rule, they correctly apply *modus ponens*, etc. On the broad reading, the experimenters are rational if their cognitive competence matches the normative standards of reasoning.

So consider the narrow reading of (1):

> (1) [narrow construal] For the irrationality experiments to prove any conclusion, the experimenters who perform them must be rational in the inferences they make by way of reaching their conclusions.

On this construal, 1 is true—clearly if the experimental conclusions rest on fallacious inferences, they do not establish any conclusions, ditto for (1), the contrapositive of this. With (1) interpreted in this way, the *reductio* no longer goes through. (3) says that since the experimenters are human, it follows from (2) that they are irrational in the sense that their cognitive competence diverges from the normative standards of reasoning. But this claim and (1) narrowly construed do not entail (4), that the irrationality experiments cannot prove any conclusion, because it is perfectly consistent with it being the case that the experimenters have cognitive competences that diverge from the norms but that, in the process of the experiments, they in fact follow the norms.

But what about the broad construal of (1), namely the claim that for the irrationality experiments to prove anything, their cognitive competence must match the normative standards of reasoning? This construal of (1) is just false. An argument that I make can be valid even if I *sometimes* make invalid arguments. Even if experimenters would violate the conjunction rule of some occasion (which no doubt they would), they may well not do so in the course of collecting and interpreting their experimental data. The *reductio* argument thus fails on either construal of premise (1).

This should not be surprising; if the *reductio* did work, it would prove too much. Consider how the argument about the irrationality experiments might be applied to visual illusions. An experimenter studying the bent-stick-in-water illusion must, of course, use vision to note the results of her experiments. It would be ridiculous to claim that since the experimenter's conclusion is that people are subject to visual illusions, all visual data is thereby called into question. Visual illusions occur only in certain contexts; their appearance does not undermine vision in general. The same sort of claim is true with respect to the irrationality experiments.

There is a further, and perhaps more important, point to be made here. Behind the *reductio* argument seems to be the suggestion that we can never move beyond our cognitive competence, that is, if we lack some heuristic in our cognitive competence, we can never acquire it: for example, if human cognitive competence lacks the conjunction rule, humans could never learn to follow it. But, at least without further argument, this does not follow. To see this, consider the analogy with linguistics and linguistic competence. It follows from linguistic theory (as articulated by Chomsky) that there are *conceivable* non-human languages that do not share some of the features all possible *human* languages share. A human child, brought up among beings (call them Martians) who spoke such a non-human language (call it Martianese), would not be able to acquire the language of her adoptive parents with the same remarkable speed with which human children, brought up by humans, normally acquire human languages.[36] This does not mean humans would be unable to learn Martianese eventually; there is no reason to think humans would never be

[36] Hilary Putnam, "The 'Innateness Hypothesis' and Explanatory Models in Linguistics," *Synthese* 17 (1967), pp. 12-22; reprinted in Ned Block, *Readings in the Philosophy of Psychology*, volume two (Cambridge: Harvard University Press, 1981), pp. 292-299, makes the stronger (and, I think, mistaken) claim that Chomsky is committed to the view that humans would be *unable* to learn Martianese (p. 292, reprinted edition).

able to learn Martianese, communicate with Martians, etc. The crucial difference for humans, between learning Martianese and English, is that we have a specialized capacity for learning languages like English but not for learning languages like Martianese. This does not, however, mean we could not learn Martianese—after all, we learn to do lots of things for which we have no specialized competence.

The relevant point with respect to reasoning is that even if humans lack some heuristic in our cognitive competence, some humans may be able to acquire it and be able to use it on some occasions. So, contra the *reductio* argument, even if the irrationality thesis is right and we lack certain rational heuristics (for example, the conjunction rule), it does not necessarily follow that certain people will not be able to learn such rational heuristics and, when they are being very careful and reflective, that they will be able to follow these heuristics. For this further reason, the *reductio* argument fails.

Earlier in this section, I noted that the irrationality thesis does not make the claim that every inference humans make are irrational. For a moment, consider this claim—what I call the *maximal* irrationality thesis—and whether a *reductio*-style argument would work against it. A plausible reaction to the maximal irrationality thesis is that the claim "every inference humans make is irrational" is self-refuting, and hence, necessarily false because it implicitly attempts to give an argument for the claim that humans cannot make rational arguments. Self-refutation is not, however, as simple a notion as it is commonly thought, and it needs to be examined closely in order to see in what way the maximal irrationality thesis is self-refuting.

J. L. Mackie usefully distinguishes three different senses in which a statement can be self-refuting.[37] First, a statement can be *pragmatically* self-refuting. Consider the statement "George cannot speak" spoken by George. The statement may well be true, for all we know, and the statement can in fact be spoken by George, but the statement cannot

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

be true *if* in fact George speaks it. This is a paradigmatic case of pragmatic self-refutation: the statement is only refuted if it is asserted in a certain way. Stricuy speaking, however, the *statement* "George cannot speak" is not self-refuting, but the action of uttering it is (the statement is contradicted by George's *act* of speaking it); that is the fingerprint of pragmatic self-refutation.[38] Second, a statement can be *absolutely* self-refuting. Consider the statement "I know that I know nothing." The statement cannot be true, because the truth of the whole sentence contradicts the claim that the sentence makes. Under no circumstances, under no way of putting forth the statement, can the statement be consistently asserted. Third, a statement can be *operationally* self-refuting. Consider th. ,entence "I believe nothing." Since saying "I believe nothing" entails that I believe that I believe nothing, the statement is self-refuting. But unlike statements that are absolutely self-refuting, operationally self-refuting cases, although they cannot be consistently put forth, could well be true; it might well be true that I believe nothing. Unlike pragmatically self-refuting statements, in which the way that the statement is put forth conflicts with the statement—George cannot say that he cannot speak, but he could use sign language to communicate it—there is no way at all that the exact sentence can be put forth—even if I use sign language, I cannot consistently assert that I have no beliefs.

With these conceptual tools, I return to the maximal irrationality thesis. Can I assert that every inference humans make is irrational? Can cognitive science discover that every inference humans make is irrational? These statements are not pragmatically self-refuting: their mode of presentation does not affect their consistency. If I put forth the statement "Every inference humans make is irrational" in some other was than speaking it, I will still be contradicting myself; even my *believing* the maximal irrationality thesis would be self-refuting. The maximal irrationality thesis is not, however, absolutely self-refuting; it could

[38] My favorite contemporary example of pragmatic self-refutation is a T-shirt popular among lesbians and gay men that says "Nobody knows I'm gay/a lesbian." It might be true of some particular person that no one knows she is a lesbian, but her wearing a T-shirt proclaiming her sexual orientation refutes the shirt's assertion.

perfectly well be the case that every inference humans make is irrational—the *statement* is not self-contradictory, even though any way that I might put it forth would be. So, the maximal irrationality thesis is *operationally* self-refuting—my asserting the statement entails that I have reason to believe it, but if the statement is true, I cannot, so the statement itself says, have any reason to believe it (or anything else for that matter). The upshot is that although the maximal irrationality thesis could not be proven by cognitive scientists or rationally believed by any human being, it still could be true, and it could be discovered, asserted and rationally believed by non-humans, for example, Martians. So while the maximal irrationality thesis is strictly speaking a limit on cognitive science, it is not a limit on the way the world is.

VII.

In the first part of this chapter, I articulated a notion of cognitive competence. The idea of a cognitive competence, based on an analogy with linguistic competence, is that humans have an underlying capacity to reason, a capacity that can manifest itself in human behavior, but that is often marred by interferences that cause performance errors. This notion is particularly crucial to the rationality thesis. Humans make mistakes in reasoning; friends of the rationality thesis do not wish to deny this fact. What they wish to deny is that the mistakes humans make are *principled* mistakes. The notion of a cognitive competence is just what they need to support the distinction between principled errors and unprincipled ones.

In the latter part of this chapter, I argued that the articulation of the notion of cognitive competence is not enough to establish the rationality thesis. In particular, I considered and criticized arguments that use the notion of cognitive competence as constituting evidence for the rationality thesis. It should not be surprising that cognitive competence alone does not establish the rationality thesis, since the very same notion will be needed by friends of the irrationality thesis as well. Suppose I think humans are irrational in that they make systematic, principled errors in reasoning. Any plausible theory compatible with this view would need to posit a cognitive competence that involves *irrational* cognitive heuristics. Even people who think humans follow irrational heuristics will allow that sometimes interferences cause humans to diverge from these heuristics. Thus, so far as this chapter is concerned, the notion of cognitive competence, while important, does not count in favor of either the rationality or the irrationality thesis.

Chapter Three: Reflections on Reflective Equilibrium

I.

Consider the distinction between normative standards and actual practice. In ethics, for example, there is a distinction between what people ought to do and what people in fact do. The same sort of distinction is applicable with respect to reasoning, namely between how people ought to reason and how tney in fact reason. That the normative-descriptive distinction is a good one does not entail that norms and practices always diverge. A truly just person would be one who always does what she ought (according to the norms of morality) to do. Similarly, a truly rational person would be one who always reasons as she ought (according to the norms of rationality) to reason. The rationality thesis is the claim that, performance errors aside, everyone reasons as she ought to. In terms of cognitive competence, the rationality thesis says a descriptive account of our cognitive competence will be identical to a normative one. The descriptive account has to do with actual human cognitive competence, that is, what rules appropriately describe human cognitive competence as it is in fact manifested, while the normative account has to do with the ideal human cognitive competence, that is, if one could have *any* cognitive competence, what would be the best to have?

The argument of central interest in this chapter, due to L. J. Cohen, is supposed to establish the rationality thesis by showing that our normative theory of rationality and our actual reasoning abilities coincide, namely that humans possess the right cognitive competence.[1] If this argument is right, then the irrationality experiments do not count in favor of the irrationalit' thesis. The argument is similar to the first argument of John Macnamara's discussed in Chapter Two, but Cohen, unlike Macnamara, provides a detailed account of how human intuitions about what counts as good reasoning play a role in the justification of the norms of reasoning. For this reason, Cohen can be seen as

[1] L. Jonathan Cohen, "Can Human Irrationality Be Experimentally Demonstrated?," *Behavioral and Brain Sciences* 4 (1981), pp. 317-370; "On the Psychology of Prediction: Whose Is the Fallacy?," *Cognition* 7 (1979), pp. 385-407; and *The Dialogue of Reason* (Oxford: Oxford University Press, 1986), pp. 149-192.

providing an answer to the objections I raised against Macnamara in section V of Chapter Two. Simply put, Cohen's argument is that the methods and data used (by cognitive scientists) to develop a characterization of human reasoning capacities are identical to the methods and data used (by philosophers) to discover and justify the norms of rationality. Since the techniques and subject matters of these inquiries are the same, the results will ultimately be the same; namely, the way we actually reason and the way we ought to reason will be the same. Though I will argue that it is unsuccessful, I think this argument for the rationality thesis is innovative and worth examining. It turns on an analogy with the competence-performance distinction in linguistics, a particular account of how cognitive psychology should be done, and a particular epistemological theory (namely, reflective equilibrium).

Drawing on the notion of cognitive competence (as discussed in Chapter Two above), Cohen compares how a descriptive theory of cognitive competence is developed to how normative principles of reasoning are developed. With respect to the latter, Cohen endorses a version of the theory of reflective equilibrium that says norms of reasoning are justified by balancing norms and actual reasoning behavior: a set of norms is justified if it fits with inferential practice and a rule of inference is accepted if it conforms to norms of reasoning.[2] Cohen goes on to describe how cognitive competence is empirically studied; on his model, an account of cognitive competence is developed in the same way the normative rules are justified. This shows the rationality thesis is true and that the irrationality experiments are either wrong or are misinterpreted when they are construed as counting in favor of the irrationality thesis.

In the sections that follow, I will examine this argument at length. In section II, I discuss the structure of the reflective equilibrium argument at greater length, contrasting it with the similar argument discussed in Chapter Two. In section III, I examine the

[2] Nelson Goodman, *Fact, Fiction and Forecast*, fourth edition (Cambridge: Harvard University Press, 1983), pp. 62-66. The term 'reflective equilibrium' comes from John Rawls, *A Theory of Justice* (Cambridge: Harvard University Press, 1971), pp. 20-21.

account—implicit in the reflective equilibrium argument for the rationality thesis—of how a psychological theory of cognitive competence should be developed. I will argue that this account is rather impoverished. In section IV, I examine the reflective equilibrium account of how norms are justified, consider an objection to it, and assess several responses to this objection. In section V, I argue that even if Cohen's account of how norms are justified can be salvaged in face of this objection, there is no isomorphism between the justification of norms and the development of a theory of cognitive competence. The reflective equilibrium argument for the rationality thesis, which is based on such isomorphism, thus fails.

II.

The reflective equilibrium argument for the rationality thesis and the first of Macnamara's arguments discussed in the previous chapter are, as I read them, variations on a straightforward and *prima facie* plausible argument for the rationality thesis: the idea is to show 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 follows:

(1) The normative standards of reasoning are based on our intuitions about what constitutes good reasoning.
(2) Our intuitions about what constitutes good reasoning come from our cognitive competence.
(3) Therefore, since the norms of reasoning come from our cognitive competence, cognitive competence must match the normative standards of reasoning.

Versions of this argument have been defended by philosophers and psychologists alike and, especially in this general form, there is something appealing about it, particularly when reasoning is compared, as it so often is by friends of the rationality thesis, to language. The comparison with language involves the fact 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. If the analogy with language is a strong one, then the argument for the rationality thesis is in good shape. In the last chapter, I argued that the analogy between language and reasoning is not strong enough to solely support the argument from (1) and (2) to (3) because linguistic norms (principles of grammaticality) are relative to linguistic competence, while norms of reasoning (principles of rationality) are *not* relative to cognitive competence. Even if the rationality thesis is true and thus human cognitive competence matches the norms of reasoning, it does not do so because the norms are indexed to actual competence in the way that the norms are with respect to language. This counts against (3) because it says that the normative standards of reasoning *are* indexed to cognitive competence.

The general argument for the rationality thesis need not, however, turn on the analogy with linguistics; one can try to make premises (1) and (2) plausible on other grounds. Cohen attempts to do so using the idea of reflective equilibrium, an account of justification on which a set of rules are justified in some domain if they provide a coherent and explicit characterization ɔf our judgments about that domain. If premises (1) and (2) can be defended independent of the failed analogy to linguistics, then the argument for the rationality thesis is immune to the objection I raised against Macnamara's version of the argument in the previous chapter. The reflective equilibrium version of the argument for the rationality thesis, which attempts to meet this challenge, is 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 is developed by a process of reflective equilibrium with our intuitions about what constitutes good reasoning as input.
>
> (RE3) Therefore, since both are developed from the same process with the same inputs, cognitive competence must match the normative standards of reasoning.

I will consider each part of this argument in turn.

III.

I first turn to (RE2), the reflective equilibrium account of how human cognitive competence should be researched.[3] 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 since 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 peopie piay chess. You wouid 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 you can test by further observing chess games. Since you do not know the rules of chess when you start, you will not initially be able to determine if some one 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

[3] Although, as I mentioned above, reflective equilibrium is a theory of how norms are justified, I shall refer to the account of how a descriptive theory of cognitive competence is developed that is implicit in (RE2) as the reflective equilibrium account of how cognitive competence is studied.

you observed some one 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, some one made the same sort of illegal move again). The rule "rooks can move vertically and horizontally unless it is September 3, 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 seem, 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.[4]

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, 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

[4] Amazingly, this is roughly how Jose R. Capablanca, the great chess master, allegedly learned how to play chess. According to his book *My Chess Career* (New York: Dover, 1965), when he was five years old, Capablanca 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. Thanks to Paul Snowden for bringing this story to my attention.

tired. A person would not agree to this rule because she believes that p implies q and p 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, etc. This strategy is supposed to be identical to the reflective equilibrium strategy used to justify norms of rationality.

Is (RE2), the reflective equilibrium account of how cognitive competence should be studied, correct? The answer, I shall argue, is no. (RE2) offers a theory of how cognitive psychology ought to be done, in particular, a theory of how a characterization of people's underlying ability to reason ought to proceed. There are two general reasons why this account is wrong: the account gives too much weight to either naive intuitions or to introspection and it ignores other sorts of data relevant to developing a theory of cognitive competence.

First, there is the issue of why people's intuitions about their own cognitive competence should be given the central role the reflective equilibrium account gives them. By "intuition," Cohen means "an immediate and untutored inclination, without evidence of inference,"[5] to make a particular judgment. He thinks this is a plausible account because of an analogy to linguistics. Cohen's insistence that naive intuitions are the ones relevant to developing a theory of cognitive competence seems a mistake for two reasons. First, the analogy with linguistic intuitions does not do what he wants it to since linguists do in fact focus on *considered* intuitions. Consider the naive intuition that

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

is ungrammatical; only on reflection is it clear that the sentence is grammatical. But it is this considered linguistic intuition that is relevant to developing an account of linguistic

[5] "Can Human Irrationality Be Experimentally Demonstrated?," p. 318.

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. Second, independent of this disanalogy, the focus on naive intuitions inevitably lets in to cognitive competence 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.

These two considerations suggest that, to be successful, the reflective equilibrium argument ought to focus on considered intuitions. 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 is developed by a process of reflective equilibrium with our *considered* intuitions about what constitutes good reasoning as input.[6]

This modification moves the reflective equilibrium argument towards co.isidering 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 pri:ciples 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 an method of research with a long tradition in psychology. It is a method, however, from which Cohen explicitly distances himself.[7] Perhaps Cohen's refusal is made for good reason since the move to introspection is problematic.

In 1879, when Wilhem Wundt set up the first psychology laboratory, introspection was *the* method of research. Subjects in Wundt's laboratory were trained to report their own

[6] The emphasis on considered intuitions is present in Macnamara, *Border Dispute*, but without any talk of reflective equilibrium.

[7] "Can Human Irrationality Be Experimentally Demonstrated?," p. 318.

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.[8] One classic experiment of this sort, performed by Saul Sternberg, 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.[9] 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 on. Further, she might continue searching through the list until the end, examining seven, nine and six asking if each is a two, even though the number two has already been found to be on the list.[10] 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

[8] For a discussion of introspection in the history and pre-history of cognitive science, see Owen Flanagan, *The Science of the Mind* (Cambridge: MIT Press, 1984). For a detailed philosophical discussion of introspection in psychology, see William Lyons, *The Disappearance of Introspection* (Cambridge: MIT Press, 1986).

[9] Saul Sternberg, "High-speed Scanning in Human Memory," *Science* 153 (1966), pp. 652-654.

[10] Sternberg's results do not in fact indicate whether subjects look through the lists from left to right, from right to left, or, as implausible as it may seem, in some other (perhaps random) order; all 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.

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 non-linguists 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 unambiguously interpret this sentence 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 actually 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, intuition (naive or considered) and introspection that psychologists can 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. If such research is performed, it would surely be relevant to a theory of cognitive competence. That, for example, some particular set of neurons is responsible for *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 has argued that basic neuroscientific facts like 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.[11] He is at pains to point out how, 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 either the (RE2) or the (RE2') account of how a theory of cognitive competence is developed.

Evolutionary psychologists such as Leda Cosmides have used evolutionary theory to help develop a characterization of human cognitive competence. In her paper "The Logic of Selection," Cosmides offers an account of the reasoning involved in the Wason selection task influenced by the constraints that evolution and natural selection place on psychological mechanisms.[12] Cosmides 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.[13]

The point is that since certain adaptive problems are likely to be important for survival and, hence, to be selectively advantageous, certain cognitive heuristics will be more likely to

[11] Christopher Cherniak, "Undebuggability and Cognitive Science," *Communications of the Association for Computing Machinery* 31 (1988), pp. 402-412. See also Cherniak, "The Bounded Brain: Toward Quantitative Neuroanatomy," *Journal of Cognitive Neuroscience* 2 (1990), pp. 58-68.

[12] Leda Cosmides, "The Logic of Selection: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task," *Cognition* 31 (1989), pp. 187-276.

[13] *Ibid.*, p. 189.

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 how we study actual cognitive competence has no room for such considerations.

Finally, Cherniak argues that computational theory is relevant to developing a theory of cognitive competence.[14] 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 criticizes Quine's "web of belief" model of human memory[15] as requiring particular heuristics—that is, heuristics for belief modification, heuristics for testing consistency, etc.—that are far too computationally demanding for the human brain to handle in a reasonable amount of time.[16] Computational considerations thus provide constraints on developing an account of cognitive competence, but such considerations are not available to the reflective equilibrium account.

These three examples of types of data that can be relevant to develop...ǯ a theory of cognitive competence count against (RE2) or (RE2'). This conclusion should be of no surprise since (RE2) gets its initial plausibility from an analogy with linguistics, a research program that includes study of more types data than just actual linguistic behavior and linguistic intuition; neuroscience,[17] theory of computation[18] and, perhaps, evolutionary theory[19] are relevant to linguistics as well. This is not to say that linguistic behavior and

[14] Cherniak, *Minimal Rationality* (Cambridge: MIT Press, 1986).

[15] W. V. O. Quine, "Two Dogmas of Empiricism," in *From a Logical Point of View* (Cambridge: Harvard University Press, 1961), pp. 20-46; and Quine and J.S. Ulian, *The Web of Belief* (New York: Random House, 1970).

[16] *Minimal Rationality*, pp. 47-54. 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.

[17] See, for example, David Caplan and Nancy Hildebrandt, *Disorders of Syntactic Comprehension* (Cambridge: MIT Press, 1988).

[18] See, for example, Ken Wexler and P. Culicover, *Formal Principles of Language Acquisition* (Cambridge: MIT Press, 1980); and Robert Berwick and Amy Weinberg, *The Grammatical Basis of Linguistic Performance* (Cambridge: MIT Press, 1980).

[19] Steven Pinker and Paul Bloom, "Natural Language and Natural Selection," *Behavioral and Brain Sciences* 13 (December 1990), pp. 707-784, argue that our innate linguistic capacity is the result of natural selection. If they are right, as I think they are, then perhaps evolutionary theory might be used to

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, howevei, to say that these considerations are not the only evidence relevant tc the study of cognitive competence.

These two objections to the reflective equilibrium account of cognitive competence—that the account cannot explain actual advances in ιhe 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 thus both false.

IV.

Having looked at the reflective equilibrium account of how a theory of cognitive competence is developed (RE2), I will now turn to the reflective equilibrium account of how norms of reasoning are justified (RE1). On the reflective equilibrium account of justification, to justify a set of principles that characterize 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

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, *Behavioral and Brain Sciences* 13 (December 1990), pp. 727-765; also see Noam Chomsky, *Language and the Problems of Knowledge: The Managua Lectures* (Cambridge: MIT Press, 1988); Steven Jay Gould, "The Limits of Adaptation: Is Language a Spandrel of the Human Brain?" talk presented to the Cognitive Science Seminar, Center for Cognitive Science, MIT (October 1987); Richard Lewontin, "The Evolution of Cognition," in *Thinking: An Invitation to Cognitive Science*, volume three, Daniel Osherson and Edward E. Smith, eds. (Cambridge: MIT Press, 1990), pp. 229-246; and, Massimo Piattelli-Palmarini, "Evolution, Selection and Cognition," *Cognition* 31 (1989), pp. 1-44.

"virtuous" circle—rules and inferences are justified together by being brought into agreement.[20] 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.[21]

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' application of reflective equilibrium to moral theory.[22] According to Rawls, in working to develop a theory of ethics, we begin with a set of moral judgments—for example, judgments like "it is wrong to torture babies"—with an eye towards producing a set of principles—for example, principles like "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

[20] It is not altogether clear whether or to what extent Goodman thinks that if some set of rules is in reflective equilibrium, then these rules are *justified* (rather than, say, reasonably believed), though the passages I cite below suggests he is committed to this reading of reflective equilibrium. Whatever Goodman's view, Cohen's argument for the rationality thesis requires reflective equilibrium to have justificatory bite; given that Cohen's argument is my concern here, I will focus on the justificatory version of reflective equilibrium.

[21] *Fact, Fiction and Forecast*, pp. 63-64.

[22] In the remaining part of this paragraph, I paraphrase *Theory of Justice*, p. 20, but omit reference to Rawls' so-called "original position."

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 upon "principles which match our considered judgments duly pruned and adjusted."[23] He calls this *reflective* because "we know to what principles our judgments conform and the premises of their derivation"[24] and *equilibrium* "because our principles and judgments coincide."[25] 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."[26]

Cohen wants to apply the reflective equilibrium account of how norms are justified to the realm of reasoning in association with his account of how actual cognitive competence is studied to defend the rationality thesis. In section III, I argued that his account of the study of cognitive competence is mistaken. For now, I will set that criticism aside to examine whether his account of how we justify our norms of rationality is right.

To begin, consider what sorts of inferential principles would result if reflective equilibrium is applied to human reasoning. There is evidence that the principles that will be in reflective equilibrium for people are principles we have very good reason to think are *not* rational. For example, as discussed above in Chapter One, even in the face of evidence and extensive briefing to the contrary, subjects in the abstract selection task experiments refuse to admit that a card showing an odd number needs to be turned over to test the rule "If a card has a vowel on one side, then it has an even number on the other side." It is unclear whether such subjects would come to accept the (proper) rule whereby all and only cards

[23] *Ibid.*
[24] *Ibid.*
[25] *Ibid.*
[26] *Ibid.*, p. 21.

with vowels and odd numbers should be examined to test the rule in question. If they do not, then they would not come to accept rational principles in reflective equilibrium.

Stephen Stich offers another such example, 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.[27] Following this fallacy and failing to properly apply the principles of logic as do subjects in the selection task are practices that would be in reflective equilibrium for lots of 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. In sum, 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 inferential strategies because they are examples of heuristics that are *not* rational but *are* in reflective equilibrium.

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.[28] For example, if the gambler's fallacy is in reflective equilibrium for someone, then it is justified for her. This seems a doomed strategy since 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

[27] See Stephen Stich and Richard Nisbett, "Justification and the Psychology of Human Reasoning," *Philosophy of Science* 47 (1980), pp. 191-193; as well as Stich, *Fragmentation*, pp. 83-84. For other examples of principles that would be in reflective equilibrium but that do not seem rational, see "Justification and the Psychology of Human Reasoning," pp. 193-195.

[28] Stich and Nisbett, "Justification and the Psychology of Human Reasoning," pp. 197-198, refer to this response as "digging in."

"bite the bullet" strategy because, as I have pointed out in Chapter Two, section IV, 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 non-starter.

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 non-rational 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 non-rational principles of reasoning would end up being in reflective equilibrium. The idea of modifying the reflective equilibrium process to prevent non-rational 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,"[29] 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 is connected to what he thinks is the right picture of linguistics and linguistic intuitions. Cohen's insistence that naive intuitions are the ones relevant to developing a theory of cognitive competence seems a mistake. As such, Cohen's hesitance to embrace (RE1') as an attempt to save the reflective equilibrium account of norms from the objection at hand is unwarranted.

[29] "Can Human Irrationality Be Experimentally Demonstrated?," p. 318.

The analogy with linguistics on which Cohen bases his focus on naive intuitions will not do the work he wants it to since, as I explained in Chapter Two, section V, *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. Since 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, (RE1'), the suggestion that only considered intuitions are involved in this reflective equilibrium process, 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. Lots of 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; not only do subjects in the experiments make systematic errors of reasoning, they sometimes stubbornly insist, even in the face of evidence to the contrary, that they are right 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 non-rational 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 right 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 standard replies to this sort of 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.[30] I will consider each possibility in turn.

[30] Stich discusses the first two views in *Fragmentation*, pp. 83-86; presumably, he thinks that what he says against the first two views counts against the third. Stich also considers is a fourth response—what Stich calls the "Neo-Goodmanian" project, *ibid.*, pp. 86-87, the project of salvaging the attempt to justify our inferential practice even in the face of the failure of the expert and wide modifications to reflective equilibrium. Our reasoning does not proceed at random, observes the neo-Goodmanian and, however it does in fact proceed, we ought to be able to describe it. Even though all attempts to do so have failed, we should not give up the project of trying to repair the idea of reflective equilibrium in order to develop an account of what it is for an inference to be justified. Stich, however, offers some arguments to the effect that we should abandon this project. He begins by pointing out that the neo-Goodmanian project is based on our commonsense notion of justification. This commonsense notion assumes there is a single, coherent concept of justification. This is an *empirical* assumption and Stich points out there is no evidence to think it is true; in fact, he thinks our intuitive notion is "messy." Stich's second criticism of the neo-Goodmanian project comes via an extended critique of a larger epistemological project (what Stich calls analytic epistemology) that uses conceptual or linguistic analysis to decide between competing rules of justification. Stich notes that the neo-Goodmanian project and analytic epistemologies are based on commonsense notions that are likely to be culturally relative. If this is right, these projects will not be suited to decide between competing candidate cognitive processes. Suppose I am trying to decide which of two different inference patterns—the one used in my culture and the one used in some other culture—is justified. According to the analytic epistemologist, I should start with a conceptual or linguistic analysis of justification. But that conceptual analysis is a culturally-relative one; it would thus be an inappropriate way to decide between two competing cognitive processes. See *Fragmentation*, pp. 86-100. I am not at all convinced by these arguments. Following Stich's critique of the Neo-Goodmanian project is beyond the scope of this chapter. Fortunately, a detailed examination of the prospects for the neo-Goodmanian project is not necessary for a satisfactory treatment of the reflective equilibrium argument. In the section that follows, I will suppose that the reflective equilibrium account of how the norms of reasoning are established is right and assess the argument for the rationality thesis on the basis of that supposition. If my supposition is false and such a project

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 from Hilary Putnam's theory of the division of linguistic labor,[31] 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.[32] Putnam's theory of meaning claims 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.[33]

This idea is to be applied in the realm of reasoning in the following fashion: some of the principles of rationality (that 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 perfectly 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

cannot be completed, the reflective equilibrium argument fails because it does not offer an adequate theory of the justification of principles of reasoning.

[31] Hilary Putnam, "The Meaning of 'Meaning,'" in *Mind, Language and Reality: Philosophical Papers*, volume 2 (Cambridge, England: Cambridge University Press, 1975), pp. 215-271.

[32] "Justification and the Psychology of Human Reasoning," pp. 198-202.

[33] "Meaning of 'Meaning,'" p. 228.

think are justified. But this, so the response goes, is a question-begging way to justify our inferential practice.[34] 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.

Stich and Nisbett 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.[35] 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.[36] 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

[34] *Fragmentation*, p. 86.

[35] "Justification and the Psychology of Human Reasoning," p. 201.

[36] For further criticism of this modification to the expert view, see Earl Conee and Richard Feldman, "Stich and Nisbett on Justifying Inference Rules," *Philosophy of Science* 50 (1983), pp. 326-331; see also *Fragmentation*, p. 164, note 16.

seemed non-experts 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 right about the norms of reasoning, that is, if an expert might accept a non-rational 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 right 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.[37] John Rawls, who coined the term "reflective equilibrium," makes a distinction between wide and narrow reflective equilibrium.[38] *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

[37] Cohen, "Can Human Irrationality Be Experimentally Demonstrated?," p. 320, explicitly rejects this suggestion. He says that the norms of reasoning "require a *narrow*, not a *wide*, reflective equilibrium" (emphasis added); see also p. 323.

[38] John Rawls, "The Independence of Moral Theory," *Proceedings and Addresses of the American Philosophical Association* 48 (1974-1975), p. 8. The distinction is implicit in *Theory of Justice*, p. 49.

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 basic 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.[39]

As an example of wide reflective equilibrium, consider Derek Parfit's argument for utilitarianism.[40] 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 loses 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. Two primary sets of intuitions that count against interpersonal balancing are the separateness of persons and our general intuitions about compensation. The separateness of persons is the fact that people

[39] For discussion of the distinction between wide and narrow reflective equilibrium, see Norman Daniels, "Wide Reflective Equilibrium and Theory Acceptance in Ethics," *Journal of Philosophy* 76 (1979), pp. 256-282; "Wide Reflective Equilibrium and Archimedean Points" *Canadian Journal of Philosophy* 10 (1980), pp. 83-103; and "On Some Methods of Ethics and Linguistics," *Philosophical Studies* 37 (1980), pp. 21-36.

[40] Derek Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984).

are separate beings, each with their own lives to lead, and, as such, are the relevant units for moral theory. This is an objection to utilitarian theory because utilitarianism sees an important role for such *inter*personal balancing. With respect to compensation, we have the intuition, for example, that if you cause me pain by breaking my leg, then, all else being equal, there is moral pressure on you to make up for that pain by compensating me. Intuitively, it would not count as compensation if you make lots of people who are total strangers happy (perhaps by making them laugh at my leg being broken). But, according to utilitarianism, if you make ***enough*** people happy as a result of your breaking my leg, you can cancel out my pain with their happiness thereby making up for the pain you have caused me. But this contradicts the strong intuition that you need to compensate me (not simply benefit some other people) because of the pain you caused me. The intuitions around both the need to compensate the person who was harmed and the "separateness of persons" both suggest that utilitarianism will not be in narrow reflective equilibrium for us.

Parfit's response to these objections can be seen as fitting the wide reflective equilibrium model. Parfit can grant that compensation and the separateness of persons count as 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 and that interpersonal compensation is acceptable. Identity of persons, Parfit argues, is not "what matters" to moral theory. Rather, according to Parfit, psychological continuity and connectedness are "what matters." But since 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, the compensation objection fails and 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. Utilitarianism is, according to Parfit, 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' discussion of the relationship of moral theory to philosophy of mind and metaphysics,[41] 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 that *I* ought to be the one compensated for pain inflicted on me) 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 match 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 to be persuaded by theoretical arguments in favor of standard probability theory, they would see that the conjunction rule is right and that they should follow it in the Linda example and similar situations; they would also 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.**

Like the other two modifications to the reflective equilibrium account of norms—the expert view and the considered intuitions view—the wide reflective equilibrium account attempts

[41] "Independence of Moral Theory," section IV.

to prevent non-rational principles of reasoning from being deemed rational by the reflective equilibrium account.

A virtue of this account is that it makes clear why the norms of rationality are not indexed to cognitive competence. Recall that, in Chapter Two, I argued 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 do not get 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 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).[42] Stich goes on to argue that it is

[42] *Fragmentation*, p. 85.

not impossible for a person to settle on a wide reflective equilibrium that includes "some quite daffy" rules of inference.[43]

One possible reply to this argument (the third modification to the reflective equilibrium view) is the expert wide reflective equilibrium view that 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 since 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 non-expert 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 at all 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.

[43] *Ibid.*, p. 86.

On the other hand, by a "daffy" principle, Stich might mean an *objectively* daffy principle, that is, a principle that is not in fact a normative principle of reasoning (regardless of whether 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 coherentist account of truth, that is, an account on which 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. Since the process of wide reflective equilibrium arguably produces a maximally coherent set of principles, then 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 its rejection of 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 that 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. Sufficc to say that, insofar as reflective equilibrium is an interesting epistemological theory, it seems a good place to look for a theory of the justification of the norms of reasoning, particularly since 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 consideration (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. In the next section, I evaluate (RE2).

V.

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, since both come from the same process with the same inputs, cognitive competence must match the normative standards of reasoning.

In section IV, 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 whose balancing processes are relevant is narrowed, and so on, then it seems the objection that the reflective equilibrium process deems some irrational principles as justified may be answered or at least the objection may be shown to be irrelevant. In section III, 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 two-fold. 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 III 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 as the inputs to the reflective equilibrium process involved in 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 III 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 III 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 III. 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 right, 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 like 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 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 the ones the reflective equilibrium process for determining cognitive competence does.

There is, howeve. , 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 weighed 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 to 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 to be successful because it must include scientific evidence in the input to the balancing process. If this modification is made and scientific evidence is included, the input to the process involved with cognitive competence and the process involved with the norms of reasoning will be different. 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 weighed in different ways given the different goals of the two reflective equilibrium processes.

VI.

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 chapter is, in a sense, no surprise given the point I made in Chapter Two against the analogy between developing a theory of linguistic competence and developing a theory of cognitive competence. I noted that norms of grammaticality are indexed to actual facts about human psychology, neurophysiology, and the like, whereas norms of reasoning are not. Since the process of developing a descriptive theory of cognitive competence is, regardless of whether 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 has made an interesting and somewhat parallel point.[44] He argues that the analogy suggested by Rawls[45] between linguistics and ethics is mistaken. Linguistics, Daniels argues, involves *narrow* reflective equilibrium while ethics involves *wide* reflective equilibrium. While I am not completely convinced that linguistics or cognitive science are appropriately characterized as *narrow* reflective equilibrium, I think that Daniels is on the right track in pointing to the distinction. 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 right about ethics and I am right 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. In light of the results of the irrationality experiments and failing a good argument that these results should be viewed as performance errors,

[44] "Some Methods in Ethics and Linguistics."

[45] *Theory of Justice*, pp. 46-48.

however, it seems the right picture of human cognition that humans have a cognitive competence which includes heuristics that do not conform to norms of reasoning.[46]

[46] Parts of this chapter were read at the June 1991 conference (in San Francisco) of the Society for Philosophy and Psychology.

Chapter Four: A Charitable Investigation of Interpretation and Charity

I

The main argument for the irrationality thesis as I have presented it rests on the irrationality experiments. These experiments, at least on the surface, show that humans violate normative standards of reasoning. Many philosophers and psychologists have argued that this is the wrong way to interpret these experiments. The reflective equilibrium argument, as discussed in the previous chapter, does not spell out how these experiments are being mis-interpreted or how they *should* be interpreted beyond seeing their results as performance errors. The reflective equilibrium argument tries to show that there are good reasons for interpreting these results in such a way that humans do not violate any norms without specifically saying how they ought to be interpreted. In this chapter, I will be more directly concerned with how the irrationality experiments should be interpreted. I will begin by looking at a specific suggestion that particular experiments are misconstrued if they are taken as support for the irrationality thesis; perhaps subjects in the irrationality experiments are misinterpreting the tasks they are being asked to perform. If so, the errors they make may not be due to their having faulty cognitive competences but rather due to some sort of comprehension errors. In section II, I explore this way of looking at the irrationality experiments as well as other strategies for reconciling these experiments with the rationality thesis. I will discuss various difficulties with these strategies. Perhaps, however, there are general considerations for thinking that the irrationality experiments should be construed in such a way that they do not provide evidence of irrationality. In search of such considerations, I turn, in section III, to Quine's principle of charity. I will discuss whether this principle might be adapted from the realm of language translation to the realm of interpreting human cognitive heuristics so as to support the rationality thesis. In section IV, I will consider an objection to the principle of charity and a modification to it known as the principle of humanity, that is supposed to answer this objection. In section V, I will look at whether the holistic version of the principle of charity might answer some objections to the principle as well as whether the move to holism is of help to the charity argument for the rationality thesis. In section V!, I will examine some arguments in favor of the principle of charity and some objections to them. In section VII, I conclude.

II.

In an article about the conjunction experiment, Kahneman and Tversky say that "a psychological analysis should apply interpretive charity and should avoid treating genuine misunderstandings as if they were fallacies."[1] Given their conclusion that humans do not follow the conjunction rule and that this rule is a norm of reasoning, Kahneman and Tversky do not mean that humans should *never* be interpreted as being irrational; rather they mean that, before interpreting humans as irrational, researchers should consider alternative explanations of human behavior. Mary Henle,[2] Braine, Resier, and Rumain,[3] and L. J. Cohen[4] make the stronger suggestion that, in fact, many of the irrationality experiments do not establish the irrationality thesis because there is good reason to think the subjects in these experiments are misinterpreting the task set before them. Henle goes so far as to say that she has "never found errors that could unambiguously be attributed to faulty reasoning,"[5] thereby advocating what I call the misinterpretation strategy for dealing with the irrationality experiments. This strategy consists of construing the irrationality

[1] Amos Tversky and Daniel Kahneman, "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgement," *Psychological Review* 90 (October 1983), p. 304.

[2] Mary Henle, "On the Relation Between Logic and Thinking," *Psychological Review* 69 (1962), pp. 376-382; and "Foreword," in *Human Reasoning*, R. Revlin and R. E. Mayer, ed. (Washington D.C.: Winston, 1978), pp. xiii-xviii.

[3] M. D. S. Braine, B. J. Resier and B. Rumain, "Some Empirical Justification for a Theory of Natural Propositional Logic," in *The Psychology of Learning and Motivation*, volume 18, Gordon H. Bower, ed. (Orlando, FL: Academic Press, 1984), pp. 313-371; they try to attribute the results of the irrationality experiments to what they call "comprehension errors."

[4] L. Jonathan Cohen, "Can Human Irrationality Be Experimentally Demonstrated?" *Behavioral and Brain Sciences* 4 (1981), pp. 317-370. Cohen calls the irrationality experiments "cognitive illusions"; he does so to parallel visual illusions (such as the bent-stick-in-the-water illusion), the idea being that subjects are not "seeing" the experimental task because the situation in which it occurs alters their "inferential vision." Gerd Gigerenzer, "On Cognitive Illusions and Rationality," in E. Eells and T. Maruszewski, eds., *Reasoning and Rationality: Essays in Honor of L. J. Cohen* (Amsterdam: Rodopi, forthcoming), argues that both Cohen and Kahneman and Tversky misapply the visual illusions analogy—on his view, statistical reasoning should be an analogy for perception, not *vice versa.*

[5] Henle, "Foreword," p. xviii.

experiments in such a way that they do not count in favoı of the irrationality thesis by claiming that subjects are misinterpreting the task they are asked to perform or are, for some other reason, failing to apply the appropriate rule.

Consider, as an example of a particular application of the misinterpretation strategy, Cohen's analysis of the selection task. In the selection task, as discussed above in Chapter One, subjects are presented with four cards from a deck containing cards with a number on one side and a letter on another. The four cards are showing, respectively, a vowel, a consonant, an even number and an odd number. Subjects are asked which cards they will need to turn over to test the truth of the rule "if a card has a vowel on one side, then it has an even number on the other side." Although the correct solution is to turn over the vowel card and the odd-number card, very few subjects do so. Cohen analyzes this experiment as follows:

> [The] experimenters' power to generate an illusion here [namely, the illusion that the odd-number card does not need to be examined to test the rule] depends on the relative unfamiliarity and artificiality of the experiment. . . . The findings about the four-card problem [i.e., the selection task] may legitimately be said to support the view that most people manage to apply their logical competence without ever formulating it expressly at a level of generality sufficient for it to be readily applicable to wholly unfamiliar tasks. . . . [S]ubjects who reason fallaciously about the four-card problem need not be supposed to lack the correct deductive "program." The subjects merely fail to recognize the similarity of their task to those familiar issues in which they have profited by using the [proper] deductive procedure . . . As a result, either that procedure received no input or its output is deleted and the behavior of the subjects manifest a matching bias [i.e., subjects fail to follow the norm].[6]

Cohen's suggestion is that in the standard selection task, subjects, due to lack of recognition or a misinterpretation of the task, do not apply the appropriate heuristic or ignore the results of applying it because they fail to realize that the task is similar to those they can and do solve correctly. Presumably, Cohen thinks that subjects (unconsciously) call up some other heuristic that generates an incorrect answer to the selection task (for example, this heuristic might cause a subject to select just the vowel card but not the odd-

[6] "Can Human Irrationality Be Experimentally Demonstrated?," p. 324.

number card). If this is right, then the fact that subjects have made mistakes in the selection task does not show that they are irrational because their cognitive competence does not match the norms of reasoning; it merely shows that subjects make performance errors—under certain circumstances, they misinterpret or fail to recognize tasks before them and thereby fail to apply correctly a rule that is in fact part of their cognitive competence.

Cohen cites what in Chapter One I called the concrete selection task as evidence for his analysis of the standard selection task. The concrete selection task involves rules like "If I go to New York, I travel by train" and cards with a name of a city on one side and a mode of transportation on the other. Subjects correctly select cards showing 'New York' and 'car' as needing to be examined in order to test the rule mentioned above. Even though the structure of this task is isomorphic to the abstract selection task, subjects perform much better when faced with concrete tasks. According to Cohen, subjects' performance on the concrete selection task suggests that subjects have the relevant deductive rules in their cognitive competence.

Friends of the irrationality thesis do not, however, need to deny this part of Cohen's conclusion about what the concrete selection task shows. The further and contentious conclusion that might be drawn from the results of the concrete selection task is that when subjects fail to apply the correct rule in the abstract selection task they are making a performance error—they have a rule in their cognitive competence that would, if applied, produce the correct result, but this rule was not applied due to some sort of interference. A friend of the irrationality thesis would disagree with this further step. She would argue that subjects' failure to apply the proper heuristics in the abstract selection task shows that they lack a *general* heuristic for performing selection tasks; subjects may have a heuristic for the selection task but the domain of application of this heuristic may be restricted to concrete situations or subjects may have a shortcut heuristic that happens to work on concrete tasks but not in general. Support for this reading of the selection task experiment can be found in experiments performed to show that subjects do not seem to be misinterpreting the selection task and hence that their failure to apply the proper rule is not due to a performance error.[7] Particularly convincing evidence for this is that even when subjects are shown (by actually presenting such a card to them) that the '7' card might have (say) an 'E' on the other side, hence violating the rule "if a card has a vowel on one side, then it will have an even number on the other," (and thus this card would be relevant to testing the rule) they insist that they were right not to select the '7' card as one that should be turned over.[8] Even when subjects are given the correct interpretation of the task, they seem to still make irrational inferences. This surprising neglect of a rule of logic even in the face of a demonstration of it suggests that subjects are truly violating the rule, not misinterpreting the experimental task.

With respect to the selection task, the central issue between friends and foes of the irrationality thesis is what the results of the various versions of the selection task indicate about human cognitive competence. Cohen, along with other foes of the irrationality thesis favors the rational competence view:

> humans have (in their cognitive competence) a heuristic that tests hypotheses of the form "if p, then q" by checking to see that when p is true, q is true and that when q is not true, p is not true, however, certain hypotheses (for example, abstract ones), give rise to performance errors that cause them to fail to apply this rule.

In favor of the rational competence view, they cite the results of the concrete selection task. In contrast, friends of the irrationality thesis favor the irrational competence view:

> Humans do *not* have (in their cognitive competence) a heuristic that tests hypotheses of the form "if p, then q" by checking to see that when p is true, q is true and that when q is not true, p is not true, however, they have a heuristic that has the effect of deeming certain hypotheses of the form "if p, then q" as true (for example, those involving concrete cases) when p is true, q is true and that when q is not true, p is not true.

In favor of the irrational competence view, they cite both the results of the concrete and the abstract selection tasks—the former show that some hypotheses do give rise to proper

[7] The experiments are surveyed in P. C. Wason and P. N. Johnson-Laird, *Psychology of Reasoning: Structure and Content* (Cambridge: Harvard University Press, 1972), chapters 14-15.

[8] *Ibid.*, pp. 196-197. Wason and Johnson-Laird (p. 196) say that subjects' denial of error in the selection task is

> surprising and frequent. . . Denial is apparent when a subject agrees that not-q associated with p would falsify the rule, but then fails to select not-q as relevant; and similarly, when he agrees that q associated with not-p is irrelevant, but then fails to dismiss q.

testing, while the latter show that other hypotheses do not. The two accounts are compatible with the behavioral evidence but they offer quite different pictures of human cognitive competence, and hence support the opposite conclusions with respect to the rationality thesis.

Consider next the conjunction experiment. As discussed in C...pter One, after reading a description of a person, subjects in this experiment are asked to rate the likelihood of various statements about the person. Recall that in the Linda case, the options they are asked to rate include:

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

According to the conjunction rule, the probability of p must be greater than or equal to the probability of p and q. Most subjects ignore this rule and rank (3) as more probable than (2). For this experiment to count in favor of the irrationality thesis, subjects must be shown to systematically violate the conjunction rule. The mere fact that subjects rate the probability that Linda is a feminist and a bank teller as being greater than that Linda is a bank teller may not, a friend of the rationality thesis could rightly point out, establish that the relevant norm is being violated; perhaps subjects are misinterpreting the various choices they are given. If subjects are misinterpreting the option before them, then their failure to rank (2) as more probable than (3) might not count against the rationality thesis.

Kahneman and Tversky have tried a large number of variations on the conjunction experiment in an attempt to isolate possible misinterpretations on the part of subjects. One such possible misinterpretation is that subjects might be reading "Linda is a bank teller" as "Linda is a bank teller, but not a feminist." If subjects are reading "Linda is a bank teller" in this way, it might explain why they think this statement is less likely to be true than "Linda is a bank teller and a feminist"; "Linda is a bank teller but not a feminist" would not be true in any of the same instances as "Linda is a bank teller and a feminist" whereas "Linda is a bank teller" would be true in some of the same instances as "Linda is a bank teller and a feminist." To test this possible explanation, subjects were presented with the following two statements about Linda and were asked to rate their probability:

> (2') Linda is a bank teller whether or not she is active in the feminist movement.
> (3) Linda is a bank teller and is active in the feminist movement.

As before, subjects rated a conjunction as more probable than one of its conjuncts even though in this case it was made clear that (2') is in fact logically equivalent to a conjunct of (3). This shows that subjects commit the conjunction fallacy even when they have been explicitly told that "Linda is a bank teller" does not mean that Linda is a *non-feminist* bank teller.[9] This suggests that subjects are not misinterpreting the task before them, that they *are* in fact systematically violating the conjunction rule and, thus, that some other (namely, some *non*-rational) cognitive heuristic guides their reasoning.

Similar (though perhaps more troubling) results were achieved when a concrete version of the experiment was performed on a group of doctors. Subjects were given a description of a patient and asked to rate the probability of various symptoms including:

> (4) The patient has hemiparesis.
> (5) The patient had dyspnea and hemiparesis.

Most of the doctors rated (5) more probable than (4) in violation of the conjunction rule. To test whether they might have been interpreting (4) as meaning "the patient has hemiparesis only (that is, the patient does *not* have dyspnea)," doctors were asked the following question:

> In assessing the probability that the patient described had a particular symptom X, did you assume that X is the *only* symptom experienced by the patient [or that] X is *among* the symptoms experienced by the patient?[10]

All but two of the sixty-two subjects said they assumed that X was *among* the symptoms rather than that X was the *only* symptom. Since doctors were not interpreting (4) to mean that the patient has only one of the symptoms, the results of this concrete version of the

[9] "Extensional Versus Intuitive Reasoning," p. 299.
[10] *Ibid.*, p. 301.

conjunction experiment suggests that doctors are following some heuristic other than the conjunction rule, that is, an *irrational* heuristic.

The motivation behind the misinterpretation strategy is to interpret the irrationality experiments in such a way that they are consistent with the rationality thesis. There are many other strategies that can reconcile the irrationality experiments with the claim that humans are rationaı. The general form of these strategies is to provide some explanation of the results of the experiments that does not involve seeing humans as having a cognitive competence that diverges from the norms.

Another strategy that reconciles the results of the conjunction experiment with the rationality thesis is that subjects know the conjunction rule is the right rule to apply in the Linda case but somehow they fail to make the correct probability judgment in spite of this. This is not quite an instance of the misinterpretation strategy—though it is similar in structure to it—since it does not suggest that subjects have misinterpreted the task; on this suggestion, they correctly interpret the task and call up the right procedure to perform it, but somehow misapply it and thereby come up with the wrong result. This suggestion has also been ıested by Kahneman and Tversky. After being presented with the description of Linda and alternatives (2) and (3)—that Linda is a bank teller and a feminist bank teller, respectively—subjects were asked to indicate which of the two arguments for (2) and (3), respectively, they found the most convincing:

> (A1) Linda is more likely to be a bank teller than she is to be a feminist bank teller because every feminist bank teller is a bank teller, but some women bank tellers are not feminists and Linda could be one of them.
> (A2) Linda is more likely to be a feminist bank teller than she is likely to be a bank teller because she resembles an active feminist more than she resembles a bank teller.[11]

A majority of the subjects chose argument (A2) that advocates the violation of the conjunction rule. These results suggest that subjects are properly interpreting the task assigned to them but still violate the conjunction rule because they call up the wrong rule.

[11] *Ibid.*, p. 299.

These versions of the conjunction experiment suggest that subjects are interpreting the tasks in the way experimenters think they are; they count against particular versions of the misinterpretation strategy and strategies that attempt to reconcile the irrationality experiments with the rationality thesis applied to the conjunction experiments.

Gerd Gigerenzer discusses a slight variation of the conjunction experiment.[12] Subjects are shown the Linda description and told that there are a hundred people who fit it. They are then asked to indicate how many of these people are:

> (a) bank tellers
> (b) bank tellers and feminists.

When the problem is phrased in such a way that subjects are being asked to indicate frequency rather than probability, subjects' disregard for the conjunction rule largely evaporates, that is, subjects' responses to this experiment are in accord with the conjunction rule.[13] The frequency version of the conjunction experiment is supposed to count in favor of the misinterpretation strategy or some similar strategy applied to the conjunction experiment. That subjects' seemingly irrational behavior disappears when some details of the task are modified (that is, when they are asked to make frequency judgments rather than probability judgments) is supposed to show that the seemingly irrational behavior is due to the details of the experiment, *not* to any problems with human cognitive competence. Friends of the irrationality thesis could respond, however, that these results underscore the irrationality of the responses to the original conjunction experiment saying that the frequency-type experiments show subjects can understand experiments similar to the conjunction experiments and reason in accordance with the norms; but why then do they fail to do so with the standard conjunction experiments? Gigerenzer's results show that sometimes humans will follow the conjunction rule; however, Kahneman and Tversky's results that subjects are not making any obvious

[12] Gerd Gigerenzer, "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases,'" *European Review of Social Psychology* 2 (1991), pp. 83-115.

[13] Klaus Fiedler, "The Dependence of the Conjunction Fallacy on Subtle Linguistic Factors," *Psychological Research* 50 (1988), pp. 123-129.

misinterpretations still suggest that humans are irrational. Though the frequency version of the conjunction experiment suggests that humans have some version of the conjunction rule in their cognitive competence, without specific support for an account of what subjects are doing in the standard conjunction task, such results can be read as highlighting rather than erasing the results of the conjunction experiments. If these results are to count in favor of the rationality thesis and also explain how Kahneman and Tversky's results are consistent with the rationality thesis, they must explain why subjects fail to apply the conjunction rule when a problem is presented in terms of probability but not when it is presented in terms of frequency.[14]

The different explanations of the data favored by Gigerenzer, on the one hand, and Kahneman and Tversky, on the other, parallel the two ways of explaining the results of the selection task. Friends of the rationality thesis favor the following version of the rational competence view:

> Humans, in their cognitive competence, have the conjunction rule but, in certain situations, (for example, when asked to make probability judgments) they make performance errors.

In contrast, friends of the irrationality thesis favor the following version of the irrational competence view:

> While humans can, in certain situations, make judgments that match what the conjunction rule dictates, they do not have, in their cognitive competence, the general conjunction rule.

Abstracting away from the conjunction and selection task experiments, for the irrationality thesis to be true, the rules instantiated in human cognitive competence must diverge from the norms of reasoning; if the irrationality experiments are to count in favor of the irrationality thesis, they must show such a divergence. In general, then, a defender of

[14] There is another—more radical—interpretation of the frequency version of the conjunction task that is suggested by Gigerenzer. The suggestion is that Kahneman and Tversky, not their subjects, are making the mistake in the conjunction experiments. According to this reading of the conjunction task, subjects are giving the correct answer to the question about Linda because the question is really asking about frequency, not probability. At least *prima facie*, this seems an odd move since the response subjects give to the Linda example makes them susceptible to a Dutch Book and that hardly can be rational. Gigerenzer thinks, however, that he can handle this worry. A full consideration of his argument is beyond the scope of this essay.

the rationality thesis could, given any particular irrationality experiment, invoke what I call the misinterpretation strategy or some other strategy that offers an explanation of why subjects fail to apply a rule that they in fact have in their cognitive competence. She could do this by arguing (as Cohen does for the selection task and Gigerenzer does for the conjunction experiment) that subjects are not violating any norms because they have misinterpreted some aspect of the task they have been asked to perform or have, for some other reason, failed to apply or attend to the correct heuristic. As with the particular experiments, friends of the rationality thesis will favor the rational competence view, the general form of which is:

> Humans have in their cognitive competence a heuristic that matches the relevant norm of reasoning but sometimes performance errors prevent this heuristic from being applied.

In contrast, friends of the irrationality thesis will favor the irrational competence view, the general form of which is:

> While humans have in their cognitive competence a heuristic or heuristics that in certain situations matches the relevant norms of reasoning, they lack the general heuristic that matches the norm and thus will sometimes behave in ways that diverge from the norm.

Whether the misinterpretation strategy can be successfully utilized as a defense of the rationality thesis, that is, whether performance errors can account for *all* the divergences from norms of reasoning might seem an empirical issue. The picture behind seeing this as empirical is as follows: a neuroscientist can crack open a human brain and examine it to see what program is instantiated therein. On this picture, the issue between the rational and irrational competence views is empirical: if the neurological examination resulted in the discovery of a program consisting of the norms of reasoning, then the rational competence view would be true and so would the rationality thesis; if a program diverging from the norms of reasoning was discovered then the irrational competence view would be true and so would the irrationality thesis. Although there is something right about this way of looking at what is at issue between the rational and irrational competence views—there is a sense in which the issues are empirical—there is something misleading about it as well. Human cognitive competence is not written in English or some familiar programming language like LISP. If the brain were cracked open, a description of cognitive competence will not simply leap out; there would remain the problem of interpreting the neurological structures that are discovered. Daniel Dennett makes a similar point when he says that even if there were such devices as "cerebrescopes" that would display a person's conscious or subconscious thoughts, either in her natural language or the "language of thought" (a.k.a., Mentalese), there would remain the question of how the output of the cerebrescope (that is, the person's internal thoughts) ought to be translated.[15] The difficulty of deducing an account of cognitive competence from observations of inferential behavior cannot be resolved simply by learning the underlying neurological structures because there exists the same sort of difficulty deducing cognitive competence from neurological structures.

This point can be made clearer by an example involving a computer. First, imagine that you are assigned the task of figuring out what program a computer is executing simply by observing its behavior. No matter how long you observe the output of the program given various input, there will be a large number of possible programs that could be behind the input-output relations you have observed. Second, imagine that you are assigned the task of figuring out the program a computer is running simply by observing its underlying physical states, that is, the states of its chips. No matter how long you observe the transition of the computer from one physical state to the next, there will still be a large number of possible interpretations of programs that these physical states could be instantiating. Trying to discover human cognitive competence is like trying to figure out what program a computer is running; the behavior and the underlying physical structure underdetermine an account of the underlying program. Of course, the big difference between the case of the computer and human cognitive competence is that, with the

[15] Daniel Dennett, "Making Sense of Ourselves," *Philosophical Topics* 12 (1981), pp. 63-81; reprinted in *The Intentional Stance* (Cambridge: MIT Press, 1987), p. 93 (references are to the reprinted version).

computer, the answer to this question may be easily accessible by asking the programmer or the chip designer or even by reading the manual. In the case of humans, the programmer-chip designer (Mother Nature?, God?, Natural Selection?) is not around and she has not left us a manual to consult.[16]

This general problem is often discussed in terms of what the frog's eye tells the frog's brain because the underlying mechanism in frogs is fairly well understood.[17] Being able to detect flies is important for frogs since flies constitute a major source of nourishment for them. Frogs accomplish fly detection with a mechanism that detects certain-sized objects (fly-sized) travelling at certain speeds (the speed that flies typically travel) and then triggers the frog's tongue in such a way and at such a time that the triggering object (hopefully, from the frog's point of view, a fly) will be caught. This mechanism makes a fairly reliable fly detector. Sometimes, however, it will not detect flies (if they are very big or very small flies or if they are travelling very fast or very slow). Also, sometimes other objects will trigger the mechanism, for example, appropriately-sized metal pellets shot past a frog at an appropriate speed. The problem is whether to call this mechanism a fly detector or whether to call it an object-of-size-X-travelling-at-speed-y detector. The very same sort of problem exists with respect to cognitive heuristics and mechanisms—even when an actual heuristic or mechanism is identified, there remains the further problem of interpreting it.

The upshot is that even if friends and foes of the rationality thesis could agree on the neurology of the human reasoning faculty, they would still disagree on how the neurological structures ought to be interpreted, in particular, they would disagree over whether the neurological structures instantiate a cognitive competence that matches the norms of reasoning. This does not mean there are no interesting empirical questions to answer with respect to the irrationality thesis, only that answering these questions will not

[16] *Intentional Stance*, *passim*, especially "Evolution, Error and Intentionality," pp. 295-300.

[17] J. Y. Lettvin, et al., "What the Frog's Eye Tells the Frog's Brain," *Proceedings of the Institute of Radio Engineers* (1959), pp. 1940-1951.

settle the debate over the irrationality thesis. The question to ask at this point is, thus, whether there are any non-empirical considerations to bring to bear on these experiments.

It is *possible* that subjects in every successful irrationality experiment that could be run would make some misinterpretation that causes them to behave in a mere *seemingly* irrational manner. Even if this misinterpretation cannot be identified, a friend of the misinterpretation strategy can insist that there must still be some subtle, as yet undetected, misinterpretation occurring. For example, recall the modified version of the conjunction experiment in which subjects were asked to choose between two arguments—(A1), which advocates following the conjunction rule, and (A2), which advocates violating it. A friend of the rationality thesis could deny that subjects' selection of (A2) indicates that human cognitive competence lacks the general conjunction rule. Instead, she could argue, for example, that subjects misunderstand the choice between (A1) and (A2) or that the selection of (A2) is a performance error. In general, experimental evidence for the irrationality thesis can always be resisted in some such way. *Prima facie*, however, the resistance techniques offered by friends of the rationality thesis are *ad hoc* immunization strategies (as discussed in Chapter Two above). Without some special theoretical considerations that suggest otherwise, it seems unlikely that such processes are plausibly behind divergences from the norms. In particular, defenders of the rationality thesis need to offer theoretical considerations that would justify the misinterpretation strategy or other strategies that reconcile the irrationality experiments with the rationality thesis.

One such consideration that would count in favor of interpreting the results of the irrationality experiments in such a way that they do not count for the irrationality thesis is the principle of charity applied to interpreting cognitive heuristics (rather than to language translation, to which it is standardly applied). That people should, in principle, be interpreted charitably (that is, in a way that maximizes the degree to which they are rational) would be a good reason for *not* interpreting the irrationality experiments as counting in favor of the irrationality thesis. Another consideration that might count in favor of interpreting the results of the irrationality experiments as counting in favor of the irrationality thesis is the fact that humans are the result of the process of biological evolution. Since evolution is a sort of optimizing process, and if, as seems likely, human cognitive competence (or a significant part of it) is innate and the result of evolution, then the fact of evolution might provide a reason for interpreting the irrationality experiments in some other way besides as counting for the irrationality thesis. I will discuss the principle of charity argument in the remainder of this chapter and turn to the evolutionary argument in Chapter Five.

III.

W. V. O. Quine, in his seminal work *Word and Object*,[18] sketches what he calls the principle of charity,[19] a principle for guiding language translation. To quote his section on "Translating Logical Connectives" at length, Quine writes:

> [L]et us suppose that certain natives are said to accept as true certain sentences translatable in the form 'p and not-p.' Now this claim is absurd under our semantic criteria. And, not to be dogmatic about them, what criteria might one prefer? Wanton translation can make natives sound as queer as one pleases. Better translation imposes our logic upon them. . .[20]
>
> That fair translation presumes logical laws is implicit in practice even where, to speak paradoxically, no foreign language is involved. Thus when to our querying of an English sentence an English speaker answers "Yes and no," we assume that the queried sentence is meant differently in the affirmation and the negation; this rather than that he would be so silly as to affirm and deny the same thing. . . .
>
> The maxim of translation underlying all this is that assertions startlingly false on the face of them are likely to turn on hidden differences in language. This maxim is strong enough in all of us to swerve us even from the homophonic method [of translation, i.e., translating a speaker's utterance as the phonetically equivalent sentence in the language that is being translated into] . . . The common sense behind the maxim is that one's interlocutor's silliness, beyond a certain

[18] W. V. O. Quine, *Word and Object* (Cambridge: MIT Press, 1960).

[19] Quine, *ibid.*, p. 59, credits N.L. Wilson, "Substances Without Substrata," *Review of Metaphysics* 12 (1959), pp. 521-539, with coining the term "principle of charity."

[20] *Ibid.*, p. 58.

> point, is less likely than bad translation or . . . linguistic divergence.[21]

The principle of charity, as Quine is offering it, is a rule of thumb that guides translation. If I translate someone as saying something absurd, then I should scrutinize my translation, not the rationality of the person. Of course, it is sometimes appropriate to translate someone as saying something absurd. For example, when Woody Allen says that his one regret in life is that he is .10t someone else,[22] he *intends* to be saying something absurd (namely, that he could be someone other than himself)—the absurdity of the comment is what makes it funny. Certainly, there are ways to construe Allen's comment so it does not violate the law of non-contradiction; for example, one could translate him as saying that he wishes he had gone into some other line of work rather than being a comedian. Such a translation, however, is not appropriate; given the context (namely, a sentence in a humor book or a line delivered in a stand-up routine), it makes sense to interpret Allen's utterance as silly and absurd. Quine need not be read as saying otherwise; Quine should be (charitably) read as focusing on "sincere assertion."[23]

To better understand the motivation for this view, recall the computer example from the previous section. Suppose a computer is running a program and you are assigned the task of figuring out what program the computer is running on the basis of the computer's behavior. Note that attempting to do so requires assuming the computer is, for the most part, working properly. To see this, consider how difficult the task would be if you did not assume that at least some of the computer's behaviors were the ones it was supposed to do. If you took every behavior of the computer as a mistake, you would have no clue at all as to the proper function of the computer; without assuming that the computer behaves properly *some* of the time, your task would be impossible. In order to offer an interpretation of the computer's behavior, you have to assume that the computer does what it should at least some of the time. This is basically the motivation behind the principle of

[21] *Ibid.*, p. 59.
[22] Woody Allen, *Side Effects* (New York: Random House, 1975), p. 151.
[23] The phrase is from Stephen Stich, *Fragmentation of Reason* (Cambridge: MIT Press, 1990), pp. 30-32.

charity: in order to interpret someone's utterances, one has to assume that their utterances are true some of the time.

Quine's injunction about translation is, however, ambiguous. Sometimes he seems to be suggesting that a person's utterances should be interpreted charitably *unless* there is strong empirical evidence to the contrary (for example, when he says that "silliness, beyond a certain point, is *less likely* than bad translation"[24]), while at other times he seems to be suggesting that people should *never* be interpreted uncharitably (for example, when he says that our logic should be *imposed* on the people we translate[25]). Rather than engaging in Quine exegesis with the aim of determining which version of the principle of charity he holds, I will simply distinguish between a stronger and a weaker version of the principle of charity. On the weaker version, a person's (sincere) assertions should not be judged irrational *unless* there is strong evidence for doing so; on the stronger version, a person's assertions should *never* be interpreted as irrational.[26]

There is much to say about the principle of charity (some of which I will say below), but, for now, the relevant question is how this principle for translating language can be applied to the irrationality experiments and the rationality thesis. Daniel Dennett and Donald Davidson, among others, have, in a number of works, argued for applying the principle of charity to the interpretation of rules of reasoning. Basically, the idea is to think of interpreting a cognitive heuristic as analogous to translating an utterance. Proper interpretation of a person's behavior, like proper interpretation of a person's utterances, requires the rationality of that person. Davidson says that "it is a condition of having

[24] *Word and Object*, p. 58, emphasis added.

[25] *Ibid.*

[26] Paul Thagard and Richard Nisbett, "Rationality and Charity," *Philosophy of Science* 50 (1983), p. 252, distinguish between five "levels of stringency" that might be applied to the principle of charity:

(1) Do not assume *a priori* that people are irrational.
(2) Do not give any special prior favor to the interpretation that people are irrational.
(3) Do not judge people to be irrational unless you have an empirically justified account of what they are doing when they violate normative standards.
(4) Interpret people as irrational only given overwhelming evidence.
(5) Never interpret people as irrational.

My weak version of the principle of charity is more or less their (3) and my strong version is more or less their (5).

thoughts, judgments, and intentions that the basic standards of rationality have application."[27] With respect to the status of the principle of charity, he writes, "Charity is forced on us; whether we like it or not, if we want to understand others, we must count them right in most matters."[28] Daniel Dennett argues that "even creatures from another planet would share with us our beliefs in logical truths"[29] and our use of the rules of logic. This is true because the principle of charity applies insofar as the person (or system) has beliefs; if a system has beliefs (that is, if it is what Dennett calls an intentional system), then it must be rational.[30] In other words, in order to interpret a person's cognitive heuristics, you have to assume that the heuristics she follows are basically rational.

Just as there is a weak and a strong version of the principle of charity as applied to interpreting utterances, there is a weak and a strong version of the principle of charity as applied to interpreting cognitive heuristics. The weak version of the principle of charity applied to reasoning is that *unless* there is strong empirical evidence to the contrary, people and the heuristics they use should be interpreted as rational. The strong version of the principle of charity is that people and the heuristics they use should *never* be interpreted as irrational. If it is appropriate to apply the principle of charity to rules of reasoning, then the principle would justify the adoption of the misinterpretation (or some other) strategy with respect to the irrationality experiments; the weak version would justify such a strategy only in the absence of strong empirical evidence against it while the strong version would justify it regardless of the evidence.

Consider the conjunction experiment as an example. The *prima facie* plausible interpretation of the results of this experiment is, following Kahneman and Tversky, that humans violate the conjunction rule and, hence, are irrational (the irrational competence

[27] Donald Davidson, "Incoherence and Irrationality," *Dialectica* 39 (1985), p. 351.
[28] Donald Davidson, "On the Very Idea of a Conceptual Scheme," *Proceedings and Addresses of the American Philosophical Association* 47 (1974), pp. 5-20; reprinted in *Inquiries into Truth and Interpretation* (Oxford: Oxford University Press, 1984), p. 197 (reference to the reprinted version).
[29] Daniel Dennett, *Brainstorms* (Cambridge: MIT Press, 1978), p. 10.
[30] See *Intentional Stance*, especially "Making Sense of Ourselves," pp. 83-101.

view). The natural interpretation of these results is not, in this case, a charitable one; a charitable interpretation would interpret people as rational. Adopting the misinterpretation strategy or some other strategy that reconciles the irrationality experiments with the irrationality thesis would thus be justified by the principle of charity applied to reasoning because these strategies do not interpret the results as showing that humans are irrational. Simply put, charity suggests that all errors are performance errors. This underscores the fact that even the strong principle of charity is compatible with attributing mistakes to people; what the strong principle of charity is *not* compatible with is the attribution of irrational cognitive competence to people.

The argument from the principle of charity for the conclusion that the results of the conjunction experiment does not count for the irrationality thesis (and hence that these results are compatible with the rationality thesis) can be spelled out as follows:

> (C1) Humans should not be interpreted as diverging from the norms of reasoning (the principle of charity).
> (C2) On the natural interpretation of the conjunction experiment, people violate the conjunction rule (a norm of reasoning).
> (C3) The natural interpretation of the conjunction experiment is mistaken; a proper interpretation of the conjunction experiment would be one that does not show people are irrational.

The argument from (C1) and (C2) justifies the application of strategies that reconcile the results of the conjunction experiments with the rationality thesis (for example, the misinterpretation strategy). This argument can easily be generalized to an argument from the principle of charity for the conclusion that the irrationality experiments do not count for the irrationality thesis (and hence are compatible with the rationality thesis) as follows:

> (1) Humans should not be interpreted as diverging from the norms of reasoning (the principle of charity).
> (2) On the natural interpretation of the irrationality experiments, people violate some norm of reasoning.
> (3) The natural interpretation of the irrationality experiments is mistaken; a proper interpretation of these experiments would be one that does not show people are irrational.

The argument from (1) and (2) justifies the application of the misinterpretation strategy or similar strategy to any of the irrationality experiments.

Are the arguments for (C3) and (3) valid? This depends on whether the principle of charity invoked in the first premise of each argument is the strong or the weak version; for this argument to provide a successful case for the rationality thesis, the *strong* version of the principle of charity needs to be adopted. Friends of the irrationality thesis can (and some do) embrace something like the weak version of the principle of charity; they claim, however, to have strong evidence for interpreting the irrationality experiments as providing support for the irrationality thesis.[31] The weak version of premise (1) would be:

> (1') Unless there is strong evidence otherwise, humans should not be interpreted as diverging from the norms of reasoning (the weak principle of charity).

Since friends of the irrationality thesis claim that there *is* strong evidence for interpreting the irrationality experiments as showing humans to be irrational, they would accept (1') but not (1); (3) does not however follow from (1') and (2). The argument is not valid if only the weak version of charity is invoked; only the strong version of the principle of charity can provide a good argument for the rationality thesis. But is the strong version of the principle of charity true? If not, then the principle of charity, regardless of whether the weak version of the principle of charity is true, would be of no help to the rationality thesis. In the sections that follow, I will set aside the weak version of the principle of charity in order to turn my attention to whether the strong version is true.

IV.

Consider the following well-worn example that is sometimes taken to count against the principle of charity.

> Suppose Paul has just arrived at a party and asserts "The man with a martini is a philosopher." And suppose that the facts are that there is a man in plain view who is drinking water from a martini glass and that he is not a philosopher [call him Biff]. Suppose also that in fact there

[31] See, for example, the sentence from Kahneman and Tversky's "Extensional Versus Intuitive Reasoning," p. 304, that I quoted above in which they endorse the *weak* version of the principle of charity; in the same article, they provide evidence for the view that humans violate the conjunction rule.

> is only one man at the party drinking a martini, that he is a philosopher, and that he is out of sight [call him Ludwig] . . .[32]

According to the principle of charity, one should interpret Paul as asserting that Ludwig, the man out of sight, is a philosopher, because this interpretation "is simple and makes his remark true"[33] and hence interprets Paul as rational. But what one really would do, and what one ought to do, is interpret Paul as asserting that *Biff*, the man drinking water from the martini glass, is a philosopher. Although this interpretation of Paul's utterance sees him as saying something *false* (since Biff is not a philosopher), it is explicable by putting ourselves in Paul's place. I can easily imagine myself both inferring that someone drinking a clear liquid out of a martini glass was drinking a martini and mistaking someone for a philosopher (for example, perhaps because he was talking about philosophy or because he looked like someone I know who is a philosopher). I can thus imagine myself in Paul's situation thinking that the man with the martini is a philosopher and, by this, meaning to refer to Biff. In contrast, I cannot imagine a plausible story of how I could be in Paul's situation and mean to refer to Ludwig when he says that the man with the martini is a philosopher.[34]

The moral of the story is that sometimes we should interpret someone as saying something false or as being irrational if doing so is the more plausible thing to do from our point of view; with respect to interpretation, preserving empathy is more important than preserving truth. This "empathic" interpretation of Paul's comment is sanctioned by a version of the principle of charity known as the principle of humanity. Richard Grandy, who coined the phrase, describes the principle of humanity as a constraint on the translation of a person's utterances and the interpretation of a person's reasoning practice in such a

[32] Richard Grandy, "Reference, Meaning and Belief," *Journal of Philosophy* 70 (1973), p. 445.

[33] *Ibid.*

[34] Actually, this is a bit fast. I can imagine lots of such stories. For example, perhaps Paul was told in advance that the only philosopher at the party will be drinking lots of martinis, or perhaps, just outside of the party, Paul ran into a friend of his who told him that Ludwig the philosopher was in the backyard drinking a martini. We can take the above example, however, to exclude these possibilities; given that Paul has no more information about philosophers and martini-drinkers at the party than what he can see when he arrives, the interpretation I describe above that interprets Paul as referring to Biff, and thus asserting a falsehood, is the *natural* interpretation even if it is not the charitable one.

way that "the imputed pattern of relations among beliefs, desires, and the world be *as similar to our own as possible.*"[35] Simply put, whereas the principle of charity constrains translation so as to maximize the rationality of the translatee, the principle of humanity constrains translation so as to maximize agreement between the translatee and the translator (that is, the person doing the translating).[36]

With respect to reasoning, whereas the standard principle of charity says that we should always interpret another as using a rational cognitive heuristic, the principle of humanity version of the charity principle says that we should always interpret another as using the cognitive heuristic that we would use in the same situation. Often, these two strategies would result in the same interpretation because often the heuristic we would expect ourselves to use would be the rational one and *vice versa*. However, sometimes, as in the martini example, the two strategies would diverge. As another example, consider charity and humanity-type principles applied to the realm of vision. If I interpreted another using a principle like charity, I would expect that she would never succumb to optical illusions, whereas, if I used the principle of humanity, I would expect that she would succumb to such illusions in situations like those in which I do.

Supposing for the sake of argument that the martini-type example counts in favor of the humanity reading of the principle of charity. Would the principle of humanity be strong enough to count as an argument for the rationality thesis? Recall the charity argument for the rationality thesis:

> (1) Humans should not be interpreted as diverging from the norms of reasoning (the principle of charity).
> (2) On the natural interpretation of the irrationality experiments, people violate some norm of reasoning.

[35] "Reference, Meaning and Belief," p. 443, emphasis added.

[36] I should note that although the principle of humanity is typically thought of as an *alternative* to the principle of charity, (see *ibid.*, pp. 439-452, *passim*), Quine in some places seems to endorse the principle of humanity. He says that when translating,

> we project ourselves into what, from his remarks and other indications, we imagine the speaker's state of mind to have been, and then we say what, in our language, is natural and relevant for us in the state just feigned (*Word and Object*, p. 219).

As before, I will avoid questions of exegesis and focus on the principle of charity and humanity as a modification of it in order to see whether either provides a good argument for the rationality thesis.

> (3) The natural interpretation of the irrationality experiments is mistaken; a proper interpretation of these experiments would be one that does not show people are irrational.

On the humanity reading of the principle of charity, (1) would be modified as follows:

> (1") Humans should not be interpreted as diverging from the principles of reasoning that I (the person doing the interpreting) would follow.

But (1") and (2) do not entail (3). To get (3), (2) needs to be modified as follows:

> (2') On the natural interpretation of the irrationality experiments, people diverge from the principles of reasoning that I would follow.

But (2') is just false. Consider the selection task. I might well select the wrong cards to examine if I am asked to turn over only those cards necessary to test the rule "if a card has a vowel on one side, then it has an even number on the other side." Even if I do not think that I would make such a mistake given my actual experiences (for example, given that I have read Kahneman and Tversky's paper many times), if my experiences had been different (that is, if I had not read their paper), I could easily make such a mistake as selecting the vowel card and the even-number card rather than the vowel card and the odd-number card. In general, the principle of humanity does not guarantee that the results of the irrationality experiments ought to be read as consistent with the rationality experiments because we might well be irrational. So, for the purposes of finding an argument for the rationality thesis, the principle of humanity will not work; if anything will provide the basis for an argument for the rationality thesis, it will be the (strong version of the) principle of charity.

That the principle of humanity will not suffice as the basis for an argument for the rationality thesis (of the form of the argument from (1) and (2) to (3)) should not be a surprise since the principle of humanity is a particular instance of the weak version of the principle of charity and the weak version of the principle of charity also fails to suffice as the basis for an argument for the rationality thesis. Recall that the weak version of the principle of charity says that unless there is strong evidence otherwise, humans should not be interpreted as diverging from the norms of reasoning. One kind of evidence that suggests humans should be interpreted as diverging from the norms of reasoning might be that *I* do in fact diverge from the norms and/or that I might do so. If this sort of evidence is considered, the weak version of the principle of charity encompasses the principle of humanity. But, as I argued above in section III, the weak version of the principle of charity will not suffice as a basis for an argument for the rationality thesis. Thus, it should be of no surprise that the principle of humanity will also not suffice in this capacity.

I began this section with the martin' example that is often taken to count against the strong principle of charity. According to the strong principle of charity applied to translation, we ought to interpret Paul as meaning that Ludwig, the philosopher who is ought of sight, is drinking a martini because, on this interpretation, what Paul says is true. But this is not how we naturally do (or should) interpret Paul. Instead, we interpret (as we ought to) Paul as meaning that Biff, the man Paul can see who is drinking water out of a martini glass, is a philosopher. But if this is right, then the strong principle of charity is mistaken—sometimes we ought to interpret people as saying things that are false or as being irrational. This, in turn, does not bode well for either the strong principle of charity applied to reasoning or its use in an argument for the rationality thesis.

Friends of the principle of charity do not, however, take this argument to be a strong one; some of them argue that the principle of charity is immune to such objections because it is supposed to be applied holistically, that is, because the principle of charity applies not to isolated utterances or individual cognitive heuristics, but to whole systems of utterances or heuristics. I turn to this response in the next section.

V.

The principle of charity applied to language is more complex than my initial exposition of it suggested. I implicitly suggested that the interpretation of an utterance was a piecemeal procedure. But this is an oversimplification of the principle of charity. The interpretation of an utterance is a *holistic* process; "only in the context of a language does a sentence have meaning."[37] The upshot is that one cannot interpret the single utterance of a person without having an interpretation (at least an implicit one) of her language in general. The same sort of point is relevant to the application of the principle of charity to reasoning and cognitive heuristics. A particular inferential behavior or a cognitive heuristic that causes such behavior cannot be interpreted on its own but rather must be interpreted holistically, that is, as part of a collection of behaviors and/or heuristics.

The holistic version of the principle of charity provides its defenders with a powerful resource for dealing with cases like the martini example in which it is tempting to see people as making mistakes. The martini example was a problem for the principle of charity because the natural interpretation of Paul's beliefs attributed a false (and, hence, uncharitable) belief to him (namely, it attributed to him the belief that Biff is a philosopher drinking a martini when in fact Biff is neither a philosopher nor a martini drinker), while the principle of charity was supposed to interpret Paul charitably, that is, as having true beliefs. But the friend of the principle of charity can now say that this was just an overly simple picture of the principle of charity. On the holistic version of charity, one does not offer interpretations in isolation; one does so only in the context of a complete interpretation of a person's beliefs, cognitive heuristics, and the like. So, according to the holistic version of the principle of charity, what matters is not only whether Paul's beliefs about Biff's profession and his drinking habits are true, but whether Paul's whole system of beliefs are for the most part true and his cognitive heuristics for the most part rational. Seen in this way, the principle of charity does not entail a counterintuitive interpretation of Paul but is compatible with interpreting Paul's belief in the martini example in the more natural way (that is, Paul believes that Biff, not Ludwig, is the martini-drinker and the philosopher); this is because the natural interpretation of Paul's belief about the martini-drinking philosopher arguably fits with the most charitable interpretation of Paul's entire belief set and his cognitive mechanisms. In general, the holistic version of the principle of

[37] Donald Davidson, "Truth and Meaning," in *Inquiries*, p. 22.

charity is compatible with attributing false beliefs in the context of a charitable interpretation of a person's entire network of beliefs.

Having elaborated the holistic version of the principle of charity, I now turn to whether it can be used to support the charity argument for the rationality thesis. Recall that this argument is as follows:

(1) Humans should not be interpreted as diverging from the norms of reasoning (the principle of charity).
(2) On the natural interpretation of the irrationality experiments, people violate some norm of reasoning.
(3) The natural interpretation of the irrationality experiments is mistaken; a proper interpretation of these experiments would be one that does not show people are irrational.

The question is whether this argument is valid when (1) is construed as the holistic version of the principle of charity. Whether it is turns on whether the holistic version of the principle of charity is compatible with seeing human cognitive competence as containing irrational cognitive heuristics. Note, however, that a feature of the holistic account, as described above, is that it allows for local error in order to make an overall interpretation more charitable. This same feature seems like it will allow for irrational cognitive heuristics in order to make for an overall charitable interpretation. The worry is that by letting in error, holism also potentially lets in irrationality; doing so has the effect of making the holistic version of the principle of charity an inappropriate premise in an argument for the rationality thesis.

Take, for example, the abstract selection task in which subjects' behavior violates the rules of logic. The misinterpretation strategy would attribute this divergence from the norms to performance errors rather than to humans having a cognitive competence that diverges from the norms. For the principle of charity to justify the misinterpretation strategy, it must require that people always be interpreted as rational, in particular, as having only rational principles in their cognitive competence. But the holistic version of the principle of charity does not have this requirement. For example, the holistic version of charity is compatible with the fact that subjects in the abstract selection task do not follow the rules of logic because they do not have these rules in their cognitive competence, as long as this divergence from the norms is offered in the context of an account that sees humans as for the most part rational. Simply put, if, for the sake of overall charity, the principle of charity permits interpreting people as making some errors, it is in danger of permitting, also for the sake of overall charity, the interpretation of people as being irrational; but once it allows for irrationality, then it no longer provides an adequate premise for use in an argument for the rationality thesis.

A friend of both the charity argument for the rationality thesis and the holistic version of the rationality thesis might try to come up with an argument that the holistic version of the principle of charity will not permit any irrationality. In other words, the idea would be to develop a defense of the strong reading of the principle of charity—that is, the principle that humans should *never* be interpreted as irrational—that in some way encompasses holism. If this could be done, then a holistic version of the principle of charity could make an adequate premise in a charity argument for the rationality thesis. I am not at all clear that this can be done. Doing so would require two steps: first, defend the strong reading of the principle of charity and, second, argue that this reading is compatible with a holistic reading of the principle of charity. In the next section, I will turn my attention to the first step: defending the strong version of the principle of charity. I will argue that the strong reading of the principle is too strong to be defensible. The charity argument for the irrationality thesis, since it requires the strong principle of charity, will thus not go through.

VI.

Quine introduced charity as a principle for guiding translation in cases of "radical interpretation." For Quine, radical interpretation involves the

> recovery of a man's current language from his currently observed responses . . . [by someone] who, unaided by an interpreter, is out to penetrate and translate a language hitherto unknown.[38]

[38] *Word and Object*, p. 28.

Quine is not, however, clear as to whether this principle should be interpreted as an empirical or a conceptual truth. I shall argue below that the principle of charity cannot be successfully defended if it is empirical. I will thus turn my attention to arguments for the conceptual truth of the (strong) principle of charity.[39]

Just as it is an empirical fact that knowledge of Latin is useful in trying to translate a hitherto unknown Romance language, it might be an empirical fact that the principle of charity should be applied in the situations requiring radical interpretation. An empirical argument for the principle of charity would go something like this:

(4) People we understand rarely make irrational (sincere) assertions.
(5) Therefore, all people rarely make irrational (sincere) assertions.
(6) Therefore, people should not be interpreted as being irrational (the principle of charity).

But it is not at all clear that we are justified in making the initial observation about the people we understand. Even assuming the validity of inductive arguments in general and this one in particular, (4) seems to already assume (6). In order for me to translate (homophonically) my neighbors' sincere English assertions, I need to invoke the principle of charity since the epistemological situation in the case of radical translation is the same as that encountered when trying to translate homophonically the utterance of someone who speaks a seemingly phonetically identical language.[40] Attempting to empirically justify the principle of charity by appeal to the seeming rationality of those people we think we understand thus begs the question; the principle of charity is thus, if true, a conceptual truth.[41]

As a digression, it is interesting to note that Grandy does not see the principle of humanity, a modification of the principle of charity, as a conceptual truth; rather he see it as

[39] In this section, unless otherwise indicated, when I talk about the principle of charity, I mean to be talking about the *strong* principle of charity.

[40] *Ibid.*, p. 78:

> we could scorn that hypothesis [that homophonic translation provides the right translation] and devise other analytic hypotheses that would attribute unimagined views to our compatriot, while conforming to all his dispositions to verbal response to all possible stimulations.

[41] For a somewhat similar argument, see *Fragmentation*, pp. 35-36.

"a *pragmatic* constraint on translation."[42] Grandy thinks the principle is pragmatic because we need to apply it only due to our epistemological situation with respect to other people's mental states. He thinks that if we could "elicit the total belief-and-desire structure [of a person, then we would be able to] . . . use mathematical decision theory"[43] to predict her behavior. We use the principle of humanity only because we do not have access to other people's beliefs and desires. But as Stich has pointed out, even if we did know all of another person's beliefs and desires and could apply decision theory to them, we would still lack an intentional characterization of that person's beliefs, that is, we would not understand what her beliefs and desires are *about*, unless her beliefs are like ours.[44]

The sort of case Grandy seems to want us to imagine is as follows: suppose that I had a complete neurophysiological description of your brain at time t. Given a description of what inputs your brain received between t and some later time t', if I knew enough about neuroscience and had a very large, very fast computer at my disposal, I could accurately predict what your brain states would be at t'. But for all that I would know about you, I would *not* know what any of your mental states are *about*. In order to know that, I need to apply the principle of humanity, for example, I need to assume that when the visual input to your brain is a red spot, you believe that there is something red in front of you; that when the auditory input to your brain is a baby crying, you believe that there is something making crying noises nearby, and so on. So even if we were not in the epistemologically impoverished situation that we in fact are in with respect to other people's mental states, we would still need to apply the principle of humanity. Grandy's claim that the principle of humanity is needed because of our epistemological situation with respect to the mental states of others is thus mistaken.

[42] "Reference, Meaning and Belief," p. 443, emphasis added.
[43] *Ibid.*
[44] *Fragmentation*, p. 47.

Further, the same reasons that prove the principle of charity is not an empirical truth also prove that the principle of humanity is not an empirical truth. An empirical justification of the principle of humanity might go as follows:

> (H4) People we understand rarely reason in ways unlike the ways we reason.
> (H5) Therefore, all people reason like we do.
> (H6) Therefore, we should interpret people as reasoning the way we do (the principle of humanity).

But unless we have already applied the ,principle of humanity, we have no good reason for thinking the people we understand reason like we do. (H4) assumes (H6), hence this argument begs the question. Without assuming the principle of humanity, an empirical justification of this principle cannot get off the ground. This is the end of the digression.

Having argued that the (strong) principle of charity is a conceptual claim, I now turn to why we should believe it. The argument, simply put, is that we should interpret people as rational because it is constitutive of what a belief is that the holder of a belief is rational. It follows from our thinking that other people have beliefs that we think they are rational and that we should interpret them according to the principle of charity. Dennett endorses this view, saying that instances of irrationality "defy description in ordinary terms of belief and desire."[45] This is not supposed to be an empirical fact but rather it is something that is true in virtue of what beliefs are, what intentional interpretation is, and so on. He writes:

> An intentional interpretation of an agent is an exercise that attempts to make sense of the agent's acts, and when acts occur that make no sense, they cannot be straightforwardly interpreted in sense-making terms. Something must give: we allow that the agent either sort of believes this or that "for all practical purposes," or believes some falsehood which creates a context in which what had appeared to be irrational turns out to be rational after all.[46]

According to Dennett, when we cannot make sense of a person's behavior through straightforward interpretation, (that is, when "something must give") we invoke "fall-back positions,"[47] that is, interpretive stances that reconcile her behavior with it being the case that she is rational (for example, the misinterpretation strategy).

[45] "Making Sense," p. 87.
[46] *Ibid.*
[47] *Ibid.*

Dennett seems to be arguing for the following conditional:

> (7) If an agent has beliefs, the agent must be interpreted as always being rational.

If (7) is true, friends of the rationality thesis need only argue for

> (8) Humans have beliefs.

in order to support the conclusion that

> (9) Humans should be interpreted as always being rational (the strong principle of charity).

If this is a good argument for (9), then the strong principle of charity can support using the misinterpretation or a similar strategy for interpreting the irrationality experiments, thereby undermining the argument for the irrationality thesis; in terms of the charity argument for the rationality thesis, the argument from (7) and (8) supports (1), the claim that humans should not be interpreted as diverging from the norms of reasoning, which, together with (2), supports (3), which in turn supports the rationality thesis.

But what exactly is the argument for (7)? Dennett does not come out and make the argument, and neither do other friends of the principle of charity such as Quine or Davidson, but it is not clear to what extent any of these thinkers would explicitly accept (7). The clearest formulation of an argument for (7) comes from Stich, who ultimately disagrees with the argument he spells out. He writes:

> It is part of what it is to be a belief with a given intentional characterization, part of the concept of such a belief, if you will, to interact with other beliefs in a rational way—a way which more or less mirrors the laws of logic. This sort of interaction with other beliefs is a conceptually necessary condition for being the belief that not-p or for being the belief that if p then q. Thus if a belief fails to manifest the requisite interactions with other beliefs, it just does not count as the belief that not-p or the belief that if p then q.[48]

The point is that what makes a mental state count as a belief is its appropriate interaction with other mental states, especially other beliefs. So, for example, a mental state X would not count as a belief that p if X is believed at the same time as not-p is. Similarly, a mental state Y would not count as the belief that if p then q if Y is believed at the same time as both

[48] *Fragmentation*, p. 37.

p and not-q. The general point is that having beliefs requires following the rules of logic. This justifies the use of the principle of charity in a limited sense—insofar as the person whose utterance we are translating or whose heuristics we are interpreting has beliefs, we must interpret her as rational in the sense that she must follow the rules of logic.

This argument for the principle of charity, if it works, only works for translation that preserves logical consistency. So, while it might work to justify using the principle of charity to support application of the misinterpretation strategy to the selection task (because subjects, in selecting the wrong cards, are violating rules of logic), the principle of charity so justified would not, for example, work to support the application of the misinterpretation strategy to the conjunction experiment. In the conjunction experiment, subjects violate the principle that the probability of p occurring should be greater than the probability of both p and q occurring (the conjunction rule). But it is possible to violate this rule without violating the rules of logic.

The argument that the rules of logic are required for having beliefs is that the rules of logic are constitutive of having beliefs, a belief that p would not count as a belief that p if it did not interact with beliefs like not-p, if p then q, and so on, in the appropriate ways. The question is whether the rules of probability are similarly constitutive of having beliefs. If not, then the sense of rationality involved in the principle of charity is not broad enough to establish that the rationality thesis is compatible with the results of all of the irrationality experiments, in particular, with those of the conjunction experiment. Friends of the rationality thesis who want to use the principle of charity argument to undermine the empirical support the irrationality experiments give the irrationality thesis thus need to argue that having beliefs requires not just the rules of logic but also the rules of probability. In particular, what the friend of the rationality thesis needs to show is that a belief that p would not count as a belief that p if the believer of p holds that the probability of p is *less* than the probability of p and q.

I am not convinced that such an argument can be made. If it cannot be made, then the principle of charity argument ultimately fails to be of help to the rationality thesis. If it *can* be made, I suspect that the argument would focus on the connection between violating the conjunction rule and being susceptible to a Dutch Book. Recall from Chapter One, that if someone violates the conjunction rule, she is vulnerable to a Dutch Book, that is, if she violates the conjunction rule, there is a set of bets that she will accept as fair such that she will lose money no matter what the outcome of the events on which she is betting. Just as it seems reasonable to say that a person does not really believe p if her belief that p contradicts her other beliefs, it might be reasonable to say that a person does not really believe p if her belief that p interacts with her other beliefs in such a way that she is susceptible to a Dutch Book. If such an argument could be made, it would show that following the rules of probability is constitutive of beliefs; this would make the principle of charity broad enough so as to support the application of the misinterpretation strategy (or a similar strategy) to the conjunction experiment and other irrationality experiments relating to rules of probability.

Assuming that such an argument could be successfully made for the rules of probability and all the other norms of reasoning (no easy task), there is still another problem facing the argument in favor of the principle of charity. In Chapter One, I discussed the consequences of what Christopher Cherniak calls the "finitary predicament" of humans.[49] One consequence of this predicament is that a person does not have enough time to check a candidate new belief for logical compatibility with every other new belief she holds; for every new belief p that she acquires, she does not have time to check all her old beliefs to see if any of them are logically equivalent to not-p. Given this, however, it is quite possible that she will sometimes come to believe both p and not-p. Yet, if someone sometimes believes both p and not-p, it is sometimes reasonable to interpret her as believing p and not-p. But, since (7) is equivalent to the claim that anyone who fails to

[49] Christopher Cherniak, *Minimal Rationality* (Cambridge: MIT Press, 1986).

preserve logical consistency has no beliefs at all, the consequences of the human finitary predicament, together with (7), leads to the counterintuitive conclusion that no human has any beliefs.

Note that the finitary predicament is only a worry for friends of the strong principle of charity, that is, a version of the principle of charity incompatible with the attribution of *any* irrationality. A friend of a weak version of the principle of charity or a friend of a hoiistic account of charity that allows for the attribution of some irrationality for the sake of overall charitability could embrace the finitary predicament and in fact use it as an argument for the weaker version of the principle of charity. Roughly, the idea would be that humans are usually rational but, due to their finitary predicament, there are some instances where their cognitive competence must diverge from the norms. The details of this are not important at the moment; the present question is whether the strong principle of charity is compatible with the human finitary predicament.

Recall that the crucial premise in the argument for the strong principle of charity is

> (7) If an agent has beliefs, the agent must be interpreted as always being rational.

One possible (though quite radical) reaction to the combination of the human finitary predicament and (7), is to bite the bullet and accept that humans have no beliefs. The argument for this conclusion would be as follows:

> (7) If an agent has beliefs, the agent must be interpreted as always being rational.
> (10) Because of our finitary predicament, humans are not rational (for example, we cannot maintain logical consistency among our own beliefs).
> (11) Therefore, humans have no beliefs.

Dennett is willing to embrace a limited version of this conclusion—for example, in the case of insane people[50]—but he seems to want to resist it in general (he seems to imply that non-insane people do have beliefs[51]), and with good reason. If you are a friend of the rationality thesis and if you want to use the principle of charity in defense of your view,

[50] *Brainstorms*, p. 10.

[51] Though he thinks that our commonsense notion of what a belief is requires scrutiny and alteration. See especially, "Beyond Belief," in *Thought and Object*, Andrew Woodfield, ed. (Oxford: Claredon Press, 1982), pp. 1-96; reprinted in *Intentional Stance*, pp. 117-202 (with postscript, pp. 203-211).

you should avoid (11), the conclusion that humans have no beliefs. The principle of charity might support the rationality thesis by suggesting that cognitive heuristics should—perhaps using the misinterpretation strategy or a similar strategy—be interpreted as being compatible with the rationality thesis. However, the principle of charity can only do this work insofar as it is applied to beliefs, cognitive heuristics, and the standard entities of psychological explanation. If humans have no beliefs and our mental states are properly characterized in some other way, there is no telling what mental state attributions would count as charitable. The principle of charity, if it works at all, only works in the familiar territory of psychological explanation; without beliefs to charitably interpret, the principle can provide no interpretive guidance and can be of no help to the rationality thesis. This is not to say that (11) is incompatible with the rationality thesis; only that the argument for the rationality thesis based on the principle of charity will not go through if humans have no beliefs.

Another possible response to the argument that the human finitary predicament counts against using the principle of charity as a guide for interpreting the irrationality experiments involves the competence-performance distinction. According to this response, the limitations resulting from the human finitary predicament affect *performance* not competence. The fact that humans do not check candidate new beliefs against all currently held beliefs to make sure they are compatible would not, on this response, be due to the fact that humans lack rational heuristics in their cognitive competence but would rather be due to performance errors.[52]

Consider an example from linguistics that this is supposed to be parallel to this case. Imagine a sentence so long that no human could have enough time in her life to read it completely. Such a sentence, despite its length, could well be grammatical. That humans cannot make this determination is not a limit on human linguistic *competence*—humans

[52] In particular, they would be performance errors due to *psychological factors*, that is, errors due to basic facts about the human condition. See Chapter Two, section II for a discussion of different factors that cause performance errors.

may have the proper heuristics in their linguistic competence to make correct judgments about grammaticality—rather, it is a limit on the amount of time that humans live, which is clearly a performance factor. This is like a computer that is programmed to generate prime numbers; no matter how long the computer runs, there will always be more numbers that it will not have had time to generate. This does not reflect poorly on the quality of the algorithm the computer is running—that is, the computer's underlying competence—rather it reflects the time limitations facing the computer.

The problem with using this move as a response to the finitary predicament objection to the principle of charity is that this move is either unsupported or it begs the question. The principle of charity is being invoked as a reason for not interpreting the results of the irrationality experiments as showing that humans have an irrational cognitive competence; in other words, the principle is being invoked to provide support for the view that humans make certain sorts of performance errors. The finitary predicament is a challenge to the invocation of this principle. Without any reason for thinking that human cognitive competence is rational, it is *ad hoc* to respond to this challenge with the claim that all errors due to the finitary predicament are performance errors. Further, to cite (whether implicitly or explicitly) the principle of charity as a reason for thinking that the finitary predicament is not relevant to cognitive competence but is only relevant to performance is to beg the question. The point is that, without particular support for so doing, friends of the rationality thesis cannot just claim that the limitations due to the finitary predicament do not apply to cognitive competence for the same reason that they cannot, without particular support, claim that the irrationality experiments do not provide any evidence about cognitive competence. The principle of charity is supposed to provide support for interpreting the irrationality experiments as pointing merely to performance errors. Friends of the rationality thesis must find another way to defend this application of the principle of charity because invoking the competence-performance distinction in the manner suggested above simply fails.

Another possible response to the finitary predicament objection to the strong principle of charity involves weakening the rationality condition in (7)—for an agent to have beliefs, she need not be *perfectly* rational, only rational enough.[53] Cherniak tries to flesh out just how "rational enough" humans need to be to have beliefs. In a nutshell, his answer is that for a mental state to count as a belief, it must interact in some significant subset of the ways that such a belief would interact if it were held by an agent who exhibited perfect rationality. For example, rather than always eliminating inconsistencies among her beliefs, according to Cherniak, to count as rational enough, an agent need only follow the "minimal consistency condition," which is:

> If A has a particular belief-desire set, then if any inconsistencies arose in the belief set, A would *sometimes* eliminate *some* of them.[54]

Cherniak argues that "minimal" rationality—*sometimes* the agent reasons in accordance with some of the norms of reasoning—is a sufficient condition for the attribution of beliefs to an agent and for an agent to qualify for psychological explanations.

The argument for the principle of charity currently under consideration is:

(7) If an agent has beliefs, the agent must be interpreted as always being rational.
(8) Humans have beliefs.
(9) Humans should be interpreted as being rational (the principle of charity).

If, however, (7) is rejected (say because of the finitary predicament of humans), this argument fails. (7) might, however, be modified by applying the notion of minimal rationality. This would result in an argument that humans should be interpreted as *minimally* rational, as *sometimes* following the norms of reasoning. The recast argument would go as follows:

(7') If an agent has beliefs, the agent must be interpreted as being *minimally* rational.
(8) Humans have beliefs.
(9') Humans should be interpreted as being *minimally* rational (call this the minimal principle of charity).

[53] See *Minimal Rationality* and *Fragmentation*, pp. 39-43.
[54] *Minimal Rationality*, p. 16, emphasis added.

The question is whether the minimal principle of charity is strong enough to undercut the irrationality thesis in the way the principle of charity is supposed to. Recall that if the strong principle of charity is justified, it can support the application of the misinterpretation strategy (or other similar strategy) to all of the irrationality experiments; so interpreted, the irrationality experiments do not count in favor of the irrationality thesis and are compatible with the rationality thesis. But the human finitary predicament counts against the strong principle of charity. The minimal principle of charity takes this predicament into account. Can, however, the minimal principle of charity take the place of the principle of charity in an argument for the rationality thesis?

Consider again the selection task. Subjects in this experiment seem, at first glance, to be violating a rule of logic. The principle of charity justifies an interpretation of the results of the selection task on which they are consistent with the rationality thesis. It does so because interpreting someone as violating a rule of logic is interpreting them as irrational. In contrast, it seems consistent with the minimal principle of charity for such a violation to be made. An agent is minimally rational if she makes *some* but not necessarily all of the inferences a perfectly rational agent would make. Thus, reasoning as a subject does in the selection task is consistent with an agent's being minimally rational; the selection task could be one of those rational inferences she does not make.

The same is true for any irrationality experiment. The strong principle of charity is of use to friends of the rationality thesis because it counts against *any* interpretation of an agent as irrational. The minimal principle of charity, since it allows for some irrationality, fails to have this use. The results of any irrationality experiment are consistent with minimal rationality so long as there are some occasions in which humans reason in accordance with the norms of reasoning (and no friend of the irrationality thesis would deny that there are *some* such occasions). Thus, the minimal principle of charity does not support the misinterpretation strategy or any similar strategy that interprets the results of the irrationality experiments as compatible with the rationality thesis.

I began this section with an eye towards defending the strong principle of charity as a guide to interpreting the irrationality experiments. If this principle could be defensibly applied to these experiments, then the natural interpretation of them—that they show humans are irrational—could be rejected in favor of some interpretation on which humans are rational. Having shown that the argument for this principle cannot be an empirical argument, I turned my attention to a conceptual argument for this principle. The conceptual argument was, in a nutshell, that charitable interpretation (that is, interpreting a person as being rational) is required for anything to count as a belief because, for a mental state to be a belief requires that it interact with other mental states according to the principles of logic. The first problem for this argument is that, even if it works to show that humans need to be interpreted as following certain basic principles of logic, such an argument does not suffice to establish enough charity to require that humans must be interpreted as rational; for example, it does not show that humans ought to be interpreted as following the principles of probability. Even if the conceptual argument works to show that humans preserve logic consistency, this is merely a partial step towards demonstrating human rationality. The second problem for the conceptual argument for the principle of charity has to do with the finitary predicament that humans face; there is just not enough time in a human life or enough space in a human brain for humans to be completely rational and hence for the strong principle of charity to be justifiably applied to interpreting the irrationality experiments. Attempting to save this conceptual argument by relegating to performance errors the limitations on human reasoning that follow from the finitary predicament is either unsupported or—if it is supported by appeal to charity—question begging. The upshot of this section is that there is no satisfactory support for using the strong principle of charity as an argument for the rationality thesis.

VII.

This chapter considered a set of arguments designed to undercut the support that the irrationality experiments seem to provide for the irrationality thesis. Even if one of these arguments had been successful, it is not clear that this would provide a strong *positive* argument for the rationality thesis, though it would have been a good negative argument since the irrationality thesis would be left without any empirical support; this would constitute a serious challenge to the irrationality thesis since the empirical support provided by the irrationality experiments was the main reason for thinking that the irrationality thesis was true. In section II, I sketched some plausible interpretations of the results of the irrationality experiments on which they were *not* interpreted as counting as evidence for the irrationality experiments. These interpretations were shown to be empirically problematic. What is needed, I argued, are conceptual considerations that favor interpretations that fit with the rationality thesis. In the remaining sections, I explored just such conceptual considerations.

In section III, I turned to the principle of charity to see if it might be successfully imported from the realm of language translation to the interpretation of cognitive heuristics for the purpose of underwriting interpreting the results of the irrationality experiments as compatible with the rationality thesis. I argued that only the strong version of the principle of charity could do so. In section IV, I considered an objection to the principle of charity and a modification to the principle of charity, known as the principle of humanity, that takes this objection into consideration. I argued that the principle of humanity is not strong enough to provide support for the rationality thesis. In section V, I considered the implications of coupling charity with interpretive holism. While this holistic version of charity seems more plausible than the principle of charity applied piecemeal, I argued that, to the extent that holism allows for interpreting people as making errors, it also allows for interpreting people as being irrational. As such, the holistic version of the principle of charity is not strong enough to support an argument for the rationality thesis. In section

VI, I considered and rejected an argument in favor cf the strong principle of charity. This, together with the conclusions of the previous sections, leads to the conclusion that considerations of charity will not suffice to support an interpretation of the irrationality experiments as compatible with the rationality thesis. Nothing I have said in this chapter counts against any version of the principle of charity except the strong version. Weaker versions of this principle may well be appropriate guides for interpreting people's utterances and their cognitive heuristics; however, such weaker versions will not suffice to support the rationality thesis.

Chapter Five: Evolution and Truth-tropic Heuristics

I.

The arguments for the rationality thesis considered so far (with the exception of the argument based on the misinterpretation strategy) have implied that the thesis is a sort of conceptual truth. In this chapter, I will consider an argument according to which the rationality thesis is an *empirical* truth. The argument, based on evolutionary theory, is initially quite appealing and many philosophers have endorsed it (though some have done so only implicitly).[1] This evolutionary argument involves two parts. The first is the claim that evolution, through natural selection, will select for cognitive heuristics and/or mechanisms that generate true beliefs. This seems reasonable because natural selection has selected other (non-cognitive) mechanisms that help in the production of true beliefs. For example, it selected the human visual system that typically does an excellent job helping us generate true beliefs about the world (for example, that there is a tiger over there, that the lights in this room are on, etc.). Further, although the visual system does sometimes generate false beliefs, these tend to be rare, to occur in identifiable ranges of situations (such as in the dark or in situations involving optical illusions), and to be compensated for fairly easily. We should expect natural selection to have done a similarly good job selecting our cognitive system. The second part of the evolutionary argument posits a connection between the use of mechanisms that produce true beliefs and rationality. It seems quite reasonable to think a system that generates true beliefs is rational; what could be more

[1] See Daniel Dennett, "Making Sense of Ourselves," in *The Intentional Stance* (Cambridge: MIT Press, 1987), p. 96; Jerry Fodor, "Three Cheers for Propositional Attitudes," *Representations* (Cambridge: MIT Press, 1981), p. 121; Elliott Sober, "Evolution of Rationality," *Sythnese* 46 (1981), p. 98; W.V.O. Quine, "Natural Kinds," in *Ontological Relativity and Other Essays* (NY: Columbia University Press, 1969), p. 126; Karl Popper, "Evolutionary Epistemology," in *Evolutionary Theory: Paths into the Future*, J.W. Pollard, ed. (London: Wiley and Sons, 1984), p. 239; Alvin Goldman, *Epistemology and Cognition* (Cambridge: MIT Press, 1986), p. 98; William Lycan, "Epistemic Value," in *Judgement and Justification* (Cambridge: Cambridge University Press, 1988), p. 142; Ruth Millikan, "Naturalist Reflections on Knowledge," *Pacific Philosophical Quarterly* 65 (1984), p. 317; and David Papineau, *Reality and Representation* (Oxford: Basil Blackwell, 1987), pp. 77-78. Not all of these people consistently support the evolutionary argument but all have been more or less tempted by it at some time. For example, Fodor, *Psychosemantics* (Cambridge, MIT Press, 1987), now eschews Darwinian explanations in cognitive science and philosophy of mind.

rational than to infer true beliefs about the world? Simply put, the two-part evolutionary argument moves from a connection between evolution and truth on the one hand and a connection between truth and rationality on the other, to a connection between evolution and rationality. The evolutionary argument is supposed to provide a strong reason for seeing humans as rational—that humans have evolved is a good reason to believe they are rational; the experiments that are supposed to show otherwise are thus either erroneous or have been misinterpreted.

The evolutionary argument that connects evolution to rationality via truth seems quite appealing to lots of people, but only its critics have attempted to spell it out carefully. In this chapter, I will argue that the evolutionary argument fails to provide support for human rationality. My discussion will focus on the first part of the evolutionary argument—the claim that evolution will select for cognitive mechanisms that produce true beliefs. I will argue that the first objection—an objection that is in favor these days, particularly in philosophical circles—is impotent. This argument tries to establish the connection between evolution and truth by way of natural selection and optimality: evolution is driven by natural selection and natural selection will select for the optimal heuristics, namely, those heuristics that select true beliefs. I will discuss four objections to this argument: first, that natural selection is not the only force that drives evolution; second, that even when natural selection explains why we have a certain trait, it does not guarantee that the trait was selected *because* of its selective advantage; third, that even if the trait was selected for its selective advantage, the trait may not be the most optimal; and, fourth, that natural selection can only choose between available traits, and often the optimal traits are not available. I will argue that the first objection is impotent. Although other evolutionary forces exist, natural selection is the only one that can plausibly explain a complex structure such as the human reasoning faculty. The other three objections pose more serious challenges to the first step in the evolutionary argument (although the second objection can be partially answered). In the last two sections, I will consider two variations on the evolutionary argument: one replaces biological evolution with conceptual evolution in the attempt to use truth to forge the link between evolution and human rationality; the other attempts to use reproductive success to replace truth as the link between evolution and rationality. I shall argue that none of the attempted modifications prove adequate to save the evolutionary argument for human rationality.

Before beginning, I should note that the evolutionary argument with which I am primarily concerned in this chapter is supposed to provide good reason for believing that human cognitive mechanisms reliably produce true beliefs. This is different from another evolutionary argument for the less radical claim that humans have beliefs. In Chapter Four, one of the motivations for the principle of charity that I discussed was that in order to interpret a person at all, one has to assume that her belief-fixing mechanisms are basically working. One might wonder why this assumption is justified, that is, why we are justified in believing that humans have beliefs and cognitive mechanisms that fix beliefs. A quite plausible answer to this wonder is to give an evolutionary account of the fact that humans have beliefs and cognitive mechanisms. This sort of evolutionary account is to be distinguished from the evolutionary argument for the rationality thesis: the former is an evolutionary account of why humans have beliefs; the latter is an evolutionary account of why humans have *true* beliefs. It is the latter argument—the one that involves truth—that might provide an argument for the rationality thesis and that is my concern in this chapter.

II.

Recall that the evolutionary argument in its general form has two parts: the first part establishes a connection between evolution and truth and the second part establishes a connection between truth and rationality. To begin, why should we believe that there is a connection between evolution and truth? In general, having true beliefs seems a good thing in terms of reproductive fitness. If I have true beliefs about what food is nourishing for me and my offspring, about where to find food, and about how to consume it, and so on, then

I will do better at getting food than I would if I had random false beliefs about these things. While this seems a plausible defense of the claim that biological evolution favors heuristics that produce true beliefs (what I call truth-tropic heuristics), there is more structure to this argument. The basic idea is that evolution is driven by natural selection, that natural selection will select optimal heuristics, and that, in turn, these heuristics will be truth-tropic ones. This seems plausible when compared to typical evolutionary theorizing about other human systems. Consider, for example, the human visual system. The standard story of its development characterizes natural selection as the driving force of evolution. Further, the particular visual system that gets selected is chosen because it is the best at doing what visual systems are supposed to do. The evolution of cognitive heuristics, in particular, and the human inferential system, in general, are supposed to work the same way—natural selection drives their evolution and, thus, the best cognitive mechanisms are selected. From here, there remains only a small further step: truth-tropic heuristics are the best cognitive heuristics and a truth-tropic inferential system is the best inferential system. And this seems, at first glance, an indisputable step; what could be a better system for belief selection than one that selects true beliefs?

Adding the second step of the evolutionary argument—namely, that there is a connection between truth and rationality—to this argument for the connection between evolution and truth produces the following argument for the rationality thesis:

> (1) Biological evolution is caused by natural selection.
> (2) Natural selection favors optimal traits.
> (3) Optimal cognitive heuristics are those that are good at producing true beliefs.[2]
> (4) Therefore, biological evolution favors heuristics that are good at producing true beliefs.
> (5) A rational heuristic is one that is good at producing true beliefs.
> (6) Therefore, biological evolution favors rational heuristics.[3]
> (7) Humans follow heuristics which are the result of biological evolution.

[2] (1), (2) and (3) are adapted from parts of an argument sketched in Stephen Stich, *Fragmentation Of Reason* (Cambridge: MIT Press, 1990), p. 63.

[3] (4), (5) and (6) are adapted from Stephen Stich, "Could Man Be an Irrational Animal?," in *Naturalizing Epistemology*, Hilary Kornblith, ed. (Cambridge: MIT Press, 1985), pp. 256-257; and *Fragmentation*, chapter three, pp. 55-74.

> (8) Therefore, there is a strong reason to believe that humans are rational—in particular, there is a strong reason not to interpret psychological experiments as demonstrating human irrationality.

In the discussion that follows, I will focus primarily on the argument for (4), the first part of the evolutionary argument.

III.

Consider the first premise (1), the claim that evolution is caused by natural selection. Although this may seem like an obvious truth, it is false; there are a variety of forces that drive evolution including random genetic drift ("sampling error" in evolution that is introduced by the finite size of populations), variable mutation rates (the phenomenon whereby if one allele mutates into another less frequently than the other mutates into it, then, whether or not it would otherwise be favored by natural selection, the allele might dominate its competitor), laws of development (for example, the correlation of head size and body size which holds for most animals), accidents of history, macromutations (the sudden appearance of totally new traits due to large-scale mutations; also known as "hopeful monsters"[4]), migration of species (the phenomenon whereby if the immigration rate of a species is higher than that of a competitor species and its emigration rate is lower than that of its competitor, then, whether or not it would otherwise be favored by natural selection, the species would be favored over its competitor), and environmental factors (for example, as George C. Williams has pointed out, natural selection does not need to be invoked to explain why flying fish have evolved so that they return to the water after entering the air; the environmental factor known as gravity will explain the evolution of such behavior[5]).[6] The existence of this variety of forces that drive evolution threatens to

[4] Richard Goldschmidt, *The Material Basis of Evolution* (New Haven: Yale University Press, 1940).

[5] George C. Williams, *Adaptation and Natural Selection* (Princeton: Princeton University Press, 1966), pp. 11-12.

[6] For a discussion of some these non-selectionist evolutionary forces, see Steven J. Gould and Richard Lewontin, "The Spandrels of San Marcos and the Panglossian Program: A Critique of the Adaptationist Programme," *Proceedings of the Royal Society of London* 205 (1978), pp. 281-288, reprinted in

unuermine the argument for (4), the claim that biological evolution favors truth-tropic heuristics,[7] because this claim seems plausible only in light of natural selection (as indicated by the role that appeal to natural selection plays in the argument from (1) and (2) to (3)).[8]

In their article "Natural Language and Natural Selection," Steven Pinker and Paul Bloom dispute claims to the effect that the evolution of language is not the result of natural selection.[9] By the evolution of language, I mean the evolution of the innate features of language, rather than of a particular type of natural language like English. Many linguists, cognitive scientists, and philosophers think that the ability to speak and understand language is biological in nature and that this biologically-based capacity for language is a distinct mental organ, an underlying mechanism (roughly analogous to other human organs except that a mental organ involves propositional attitudes) that embodies linguistic knowledge.[10] Pinker and Bloom argue that natural selection is the only explanation of the evolution of the language faculty- -none of the other forces that drive evolution can explain

Conceptual Issues in Evolutionary Biology, Elliott Sober, ed. (Cambridge: MIT Press, 1984), pp. 252-270; see also Elliott Sober, *The Nature of Selection* (Cambridge: MIT Press, 1984), pp. 20-31.

[7] In contrast, Lycan, "Epistemic Value," p. 153, suggests that even if nonselectionist evolutionary forces are responsible for our cognitive mechanisms, the rules that our mechanisms follow would be basically the same as those we would have had if natural selection were the sole cause of them. The two very brief arguments he gives for this claim are not successful.

[8] This argument against the conclusion that biological evolution favors truth-tropic heuristics is similar to one made by Stich, *Fragmentation*, pp. 63-64. Sober, "Evolution of Rationality," pp. 110-111, expresses concern that nonselectionist forces might undermine the evolutionary argument. Richard Lewontin, "The Evolution of Cognition," in *Thinking: An Invitation to Cognitive Science*, volume three, Daniel Osherson and Edward E. Smith, eds. (Cambridge: MIT Press, 1990), pp. 229-246, also suggests that, in light of various nonselectionist forces of evolution, our cognitive mechanisms might have evolved but not be the result of natural selection.

[9] Steven Pinker and Paul Bloom, "Natural Language and Natural Selection," *Behavioral and Brain Sciences* 13 (December 1990), pp. 707-784. Among those who have claimed that the evolution of natural language is *not* due to natural selection are Noam Chomsky, *Language and the Problems of Knowledge: The Managua Lectures* (Cambridge: MIT Press, 1988); Steven Jay Gould, "The Limits of Adaptation: Is Language a Spandrel of the Human Brain?" talk presented to the Cognitive Science Seminar, Center for Cognitive Science, MIT (October 1987); Richard Lewontin, "The Evolution of Cognition"; and, for an extended discussion of the view, Massimo Piattelli-Palmarini, "Evolution, Selection and Cognition," *Cognition* 31 (1989), pp. 1-44. See also the commentaries on "Natural Language and Natural Selection," *Behavioral and Brain Sciences* 13 (December 1990), pp. 727-765.

[10] See, for example, Noam Chomsky, *Knowledge of Language: Its Nature, Origin and Use* (New York: Praeger, 1986); and Jerry Fodor, *The Modularity of Mind* (Cambridge: MIT Press, 1983). See also my discussion of mental organs and mental modules in Chapter Two, section II.

the intricate structure of the innate mental organ behind the ability to speak and understand natural languages. Their argument for this thesis involves two major steps: first, they argue that natural selection is the only candidate explanation for the evolution of functionally complex structures and, second, they argue that the language faculty is such a complex structure. I am here concerned with the first step in their argument.[11] If this step is right and if the human inferential system is, like the language faculty, a functionally complex mental organ, then (1) could be modified so as to be immune to the objection that there are other forces besides natural selection that drive the evolution of cognitive mechanisms—the existence of a variety of forces that drive evolution only counts against (4) if there is reason to believe these other forces play a role in the evolution of cognitive mechanisms. Premise (1) can, thus, be modified as follows without any loss to the argument:

> (1') The biological evolution *of functionally complex structures* (including human cognitive mechanisms) is caused by natural selection.[12]

I now turn to the argument for (1').

IV.

Williams's point that "precision, economy, efficiency, etc., . . . rule out pure chance as an adequate explanation"[13] of a trait is nicely elaborated by Pinker and Bloom:

> The essential point is that no physical process other than natural selection can explain the evolution of an organ like the eye. The reason for this is that structures that can do what the eye does are extremely low-probability arrangements of matter. By an unimaginably large margin, most objects defined by the space of biologically possible arrangements of matter cannot bring an image into focus, modulate the amount of incoming light, respond to the presence of edges and depth

[11] Pinker and Bloom draw this point from *Adaptation and Natural Selection*, see especially chapters 1, 2, and 9.

[12] This thesis has been defended by evolutionary psychologists (among others). See, for example, John Tooby, "The Emergence of Evolutionary Psychology," in *Emerging Synthesis in Science*, D. Pines, ed. (Redwood City, CA: Addison-Wesley, 1988); and Leda Cosmides and John Tooby, "From Evolution to Behavior: Evolutionary Psychology as the Missing Link," in *The Latest on the Best: Essays on Evolution and Optimality*, John Dupré, ed. (Cambridge: MIT Press, 1987). Evolutionary psychologists do not necessarily endorse (4) or the argument for it from (1'), (2) and (3).

[13] *Adaptation and Natural Selection*, p. 10.

> boundaries, and so on. The odds that genetic drift, say, would result in the fixation within a population of just those genes that would give rise to such an object are infinitesimally small, and such an event would be virtually a miracle. This is also true of the other nonselectionist mechanisms . . . Natural selection—the retention across generations of whatever small, random modifications yield improvements in vision that increase chances of survival and reproduction—is the only physical process capable of creating a functioning eye, because it is the only physical process in which the criterion of being good at seeing can play a causal role. As such, it is the only process that can lead organisms along the path in the astronomically vast space of possible bodies leading from a body with no eye to a body with a functioning eye.[14]

Having argued that natural selection is the only good explanation for the human eye, Pinker and Bloom generalize the argument: any product of evolution that shares with the human eye a functionally complex, articulated, and intricate structure—a structure "composed of many interacting parts where the details of the parts' structure and arrangement . . . fulfill some function"[15]—must be explained through natural selection because such a structure requires the existence of selective retention on the basis of "criterion of being good" at some function to explain its appearance.[16]

The point is that natural selection is the only good scientific explanation for the evolution of functionally complex structures. For this to be relevant to the evolutionary argument for rationality, one needs to show that the human inferential system is both functionally complex and the result of evolution. In what follows, I sketch reasons for believing this. First, I suggest that there is good reason to think that the human inferential system is (at least partly) innate, and, hence, the result of evolution. Second, I argue that this system is reasonably thought to be a mental organ (though not a mental module). Third, I argue that this mental organ is functionally complex. This, coupled with Williams'

[14] "Natural Language and Natural Selection," p. 710. The complexity of the human eye (as well as "radar" in bats) is eloquently discussed in Richard Dawkins, *The Blind Watchmaker* (New York: W.W. Norton, 1986).

[15] "Natural Language and Natural Selection," p. 710.

[16] One potential worry about this thesis is that it seems to require having a notion of complex design. That complex structures cannot be explained by non-selectionist forces while non-complex structures can assumes an account of what it is to be complex. On the basis of what criterion are we to distinguish the human eye as complex and some random collection of molecules as non-complex? This might be worrisome if evolutionary biologists of all stripes did not often help themselves to this very notion. Even biologist critics of some features of mainstream evolutionary theory (for example, Gould) make use of intuitive motions such as complexity and design. Pinker and Bloom's account is no more problematic in this way than those of the people who hold the view they are criticizing. See *ibid.*

point about complex structures, amounts to a strong argument that the human inferential system is the result of natural selection.

To begin, at least some basic part of the human inferential system must be innate because some ability to draw inferences is required to learn *anything*, including principles of reasoning. One cannot learn if one does not already have some skill in making inferences on the basis of experience; this is almost true by definition—one cannot learn anything without (at least implicitly) knowing the principles or having the mechanisms that enable one to learn. Even the most anti-nativist of thinkers would agree to this argument for the innateness of *some* principles of reasoning, but this argument is quite limited in scope as it only gives us reason to believe that there is some very small set of innate principles of reasoning. For the evolutionary argument to be an interesting one, we must have good reason to think that a significant number of our cognitive heuristics are innate.

The other reasons for accepting the innateness of human cognitive heuristics are less strong. Together, however, they suggest that it is *plausible* that much of our reasoning ability is innate. First, humans seem to be able to apply cognitive heuristics without having had adequate experience to learn them. For example, children seem to follow the conjunction elimination rule (if you believe A and B, then you should believe A) without ever having been explicitly taught it and without having experienced enough instances of the conjunction elimination rule being applied to reasonably induce it for themselves. This parallels the poverty of stimulus argument about language. Chomsky observes that children come to understand and speak language with such speed and sophistication as to make it impossible that they learned to do so by being explicitly taught to speak or by listening to other people speak and inferring principles of language from them. In other words, the input children have received grossly underdetermines a specific set of rules of language; simply put, children who understand and speak a language have not had the adequate stimulus to have learned it. The same sort of point seems applicable to reasoning. Children can make all sorts of inferences that they have neither been explicitly taught how to make nor do they seem to have had the occasion to adequately observe the principles being applied.

A second consideration that seems to make it plausible that our principles of reasoning are innate is that they are often unconscious. Some people, for example, apply *modus ponens* without knowing that they are doing so and without being able to articulate the principle at all. The case is even stronger with more complex principles of reasoning; such principles may well be followed even though people do not realize that they have been following them and are not able to articulate them. But does the fact that a principle is unconscious provide evidence that it is innate? In some cases it does not. A ubiquitous philosophical distinction is between *doing* and *describing*. Just because I can *do* something (for example, tie my shoes, ride a bike) does not mean that I can *describe* how it is that this task is done (for example, though I can ride a bike perfectly well, about the best I can do to describe how it is done is to say "Get on that bike, keep your balance, keep your wits about you, and just ride!"). It is quite natural to say that I am not conscious of how it is that I do something even though I can perfectly well do it. (I perhaps can do some things *better* in virtue of their being unconscious; for example, when I think about how it is that I ride a bike *while* riding one, I find it much more difficult to do and am much more likely to fall off the bike.) But all this could be true and the relevant principles might not be innate. For example, my ability to ride a bike is presumably *not* innate but learned; the explanation for my being unconscious of how I ride my bike is that I have "internalized" how to ride a bike, it has become a sort of reflex. The relevance of all this to reasoning is as follows: the fact that some principles of reasoning are unconscious, unarticulated, etc., does not entail that they are innate; they may have been learned and internalized.

Third, there is some evidence to suggest that the principles of reasoning that we follow (in this culture) are universally followed (by humans).[17] Fourth, there seems to be some

[17] Edwin Hutchins, *Culture and Inference: A Trobriand Case Study* (Cambridge: Harvard University Press, 1980).

evolutionary continuity with respect to human reasoning ability as compared to that of other species.[18] The universality and animal evidence are only suggestive of innateness. Universality could be explained in other ways and the ability to learn principles of reasoning might be what is relatively continuous between animals and humans, not any innate principles of reasoning. Fifth, there is some sketchy neurological evidence for the innateness of reasoning; certain neurological disorders seem to cause problems with reasoning.[19] As I admit above, none of this evidence *proves* that our reasoning faculty is mostly innate (especially since the evidence is sketchy), however, it suggests that it is not implausible to think that it is innate.

Having sketched several reasons for thinking that the humans have an innate reasoning capacity, I turn to an argument against this view. Stich argues that the cognitive strategies we use are not, for the most part, innate. He says that

> it *may* be the case that the strategies of reasoning a person employs, like the language he or she speaks, are determined in large measure by environmental variables and that variation in inferential strategies across persons or societies are largely independent of genetic factors.[20]

Whether this is a plausible claim depends on what exactly Stich means by "may" here. Of course, there is a sense in which it *might* be the case that most (though not all, because of the first argument for innateness given above) of our inferential principles are learned, in the same way in which it *might* be the case that people learn how to breathe or know innately how to drive a car with standard transmission. The more interesting question is whether we have any good reason for thinking that people *actually do* learn the inferential

[18] C. R. Gallistel, *The Organization of Learning* (Cambridge: MIT Press, 1990); Steven Walker, *Animal Thought* (Boston: Routledge, 1983); S. T. Parker and K. R. Gibson, *Language and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives* (Cambridge: Cambridge University Press: 1990); Howard Rachlin, A. W. Logue, John Gibbon, and Marvin Frankel, "Cognition and Behavior in Studies of Choice, *Psychological Review* 93 (1986), pp. 33-45; Richard Hernnstein, "Level of Stimulus Control: A Functional Approach," *Cognition* 37 (199), pp. 133-166; and Donald Griffin, *Animal Thinking* (Cambridge: Harvard University Press, 1984).

[19] See, for example, Alan Leslie and Uta Frith, "Prospects for a Cognitive Neuropsychology of Autism: Hobson's Choice," *Psychological Review* 97 (1990), pp. 122-131; Scott Atran, *Cognitive Foundations of Natural History* (Cambridge: Cambridge University Press, 1990); and Alfonso Caramazza, John Hart, and Rita Berndt, "Category-Specific Naming Deficit Following Cerebral Infraction," *Nature* 316 (August 1985), pp. 439-440.

[20] *Fragmentation*, p. 69, emphasis added.

practices we use. To the extent that Stich attempts to make such an argument, it is a weak one.

Stich's argument turns on an analogy with language. He observes that if he had been born in Braz:l, he would speak Portuguese, and if he had been born in Korea, he would speak Korean. Clearly, he says, which language a person speaks is not under genetic control but is rather due to a variety of social factors. Although he does not present any evidence for the strength of the analogy, Stich says that

> it might be the case that language acquisition and inferential system acquisition are parallel . . . it might well turn out that neither our entire inferential system nor any part of it is innate or universally shared among humans.[21]

I actually think that the analogy to language is a quite useful one here, but I do not think Stich draws the correct moral from it. The moral Stich wants to draw is that just as there is linguistic diversity that shows language is not innate, there is cognitive diversity that shows inferential principles are not innate. But this is not the right moral to draw. An example of the sort of evidence Stich is thinking about when he appeals to linguistic diversity is the case of two identical twins, one raised in a Korean-speaking environment and one raised in a Portuguese-speaking environment. The twins here have the same genes but speak different languages. This shows that the language people speak is not genetically controlled and hence not innate. This is of course right, but the conclusion can be pushed too far. Consider the case of two twins, one brought up in an environment where she eats a very balanced diet, gets lots of exercise, and so on, and the other brought up in an environment where she eats very poorly and gets no exercise; the twins will end up with radically different body sizes, habits, personalities, and the like. This would show, for example, that body size is not completely independent of environmental factors, but it does not show that body size is not *primarily* genetically controlled. The same point applies to language: just because different people speak different languages does not mean that

[21] *Ibid.*, p. 72.

language is not *primarily* genetically controlled. So, although there is linguistic diversity, it is the sort of diversity that is compatible with nativism about language in even the strong sense.[22]

Above, I said that the analogy between language and reasoning with respect to the degree of innateness is a good one, but not in the way Stich thinks it is. According to Chomsky and friends, there is an innate underlying universal grammar that is behind all actual and possible (given our brains) human languages. This is both consistent with the diversity of human languages and explains the similar abstract structure that all human languages have. Similarly, the view that our inferential principles are innate does not suffer from the vice of precluding *some* diversity of cognitive ability; rather, it has the virtue of explaining why humans have such structurally similar inferential principles.[23]

If the human inferential system is, as I have suggested, innate, what needs to be shown is that the human inferential system is a mental organ—namely, that it is the product of a set of computational modules in the brain—and that it is functionally complex—that is, complex in the sense that the chances are infinitesimally small that its parts fit together in non-trivial ways to do non-trivial things. To start, consider the human eye: it is made up of many parts and many genes play a role in its development. The eye, however, performs a

[22] Stich, "Empiricism, Innateness, and Linguistic Universals," *Philosophical Studies* 33 (1978), pp. 273-286, distinguishes between four different flavors of "non-empiricist theories of language acquisition" in attempt to argue against certain aspects of Chomsky's nativism. His arguments there seem to me quite unconvincing. Supporting this contention is beyond the scope of this project.

[23] Stich argues, *Fragmentation*, pp. 72-73, that even if human cognitive competence is innate, there may still be *great* variation what heuristics humans have because of genetic differences. There is good reason to think that Stich is wrong about this. As Cosmides and Tooby, "From Evolution to Behavior," p. 304, write:

> Of course there can be individual variation in cognitive programs, just as there is individual variation in the size and shape of stomachs: this can be true of any structure or process in a sexually recombining species, and such genetic variation constitutes that basis for "inherited" or "constitutional" differences. However, because even simple cognitive programs or "mental organs" must contain large number of processing steps, and so must have complex polygenic bases, they must necessarily evolve slowly, leading to variation being mostly "superficial." There is a large amount of variation among humans concerning single or quantitative characteristics of specific organ systems, but there is almost *no variation* among humans in what organs exist. Everyone has a heart and a liver, and so on, and everyone's heart and liver function in much the same way. We expect that this pattern holds for "mental organs" as well. . . . We find implausible, on the basis of population genetics considerations, the notion that different humans have fundamentally different and competing cognitive programs, resting on wholly different genetic bases. . . .

particular function and the various parts of the eye are useful insofar as they play a role in enabling the eye to perform its function; these are among the reasons for thinking of the eye as a single organ. In contrast, there are no good reasons for thinking of the parts of the body involved in hitting a baseball as being the parts of a "batting organ."[24] For example, the various parts of the body involved in hitting a baseball have their own separate uses (vision, locomotion, etc.). Consider now the claim that there is a mental organ responsible for human reasoning. Human reasoning is like the eye and the language faculty and *un*like the body parts involved in batting. Although there are various parts to the human reasoning system (that is, various heuristics involved in making inferences, probability judgments, etc.) and although these parts may result from different genes, they work together to perform an adaptive function (reasoning) and they are useful only in virtue of playing a role in the reasoning process. This suggests that there is what might be called an inferential or reasoning organ.

Note, however, that none of what I have said suggests that the inferential organ is a mental module, that is, a domain-specific, innately specified, hard-wired, autonomous mental organ. In fact, there is good reason for suspect that it is not. Recall from Chapter Two, section II that for a mental organ to be a mental module it must be domain-specific. Our principles of reasoning, unlike, for example, rules of language, are quite general, they need to be applied in a whole range of domains. This is because we can have beliefs about anything and we may need to make inferences on the basis of any of these beliefs. The generality of the principles of reasoning suggests that they are not modular.[25]

Having suggested that there is an innate reasoning organ, I turn to the claim that this organ is functionally complex. The point simply is that the chances are infinitesimally small that a random set of molecules would come together to form a structure that could perform the functions the reasoning organ performs. A good way to see this point is to

[24] The example is similar to one used in *Modularity*, p. 20.
[25] *Modularity, passim.*

think of the set of systems that are computationally equivalent to the reasoning faculty. Although systems in this set could be instantiated on all different sorts of computer architectures, they can be thought of as a set of Turing machines. A Turing machine is an abstract mathematical model of computation. There is a Turing machine equivalent of every possible computational system and there are an infinite number of Turing machine configurations, many of which do nothing of any functional significance. The odds that some randomly selected Turing machine will do anything of interest are quite small; more relevantly, the odds that a randomly selected Turing machine would work to fulfill the cognitive needs of humans are infinitesimally small. Only a process which involves the selective retention of changes (that is, changes due to feedback between a process and the environment in which the process takes place) could produce such a system with any significant degree of probability.

The human ability to reason is flexible, general and intricate. The inferential system works quickly, regularly, in a variety of domains, and, the experiments that are supposed to demonstrate human irrationality aside, consistently and reliably. These considerations suggest that there was some feedback between the evolution of the infer^nce faculty and its success at going about the business of reasoning. This sort of feedback is the fingerprint of natural selection; none of the other evolutionary forces can explain such functional complexity.

To review, I have sketched several reasons for thinking that our capacity for reasoning is innate. Some principles of reasoning are necessary for any learning to occur, we seem to apply such principles before we could have learned them from experience, the principles are often unconscious and yet they seem like they have not been internalized, the principles seem universal, the principles seem to be neurologically hard-wired, and animals seem to have some similar heuristics. Further, this capacity seems to fit the model of a mental organ, and a complex one in the sense that only natural selection could produce it. This does not at all constitute an airtight argument in favor of the claim that natural selection is the force that drove the evolution of the human capacity to reason. It does however shift that the burden of proof to those who claim some other evolutionary force is at play in the development of the inferential organ. Since natural selection is the only good explanation of complex design and since the inferential organ is complex, there are strong reasons for thinking that the evolution of cognitive mechanisms was caused by natural selection.

V.

The argument for the claim that evolution favors truth-tropic mechanisms (4) currently under consideration is:

> (1') The biological evolution *of functionally complex structures* (including human cognitive mechanisms) is caused by natural selection.
> (2) Natural selection favors optimal traits.
> (3) Optimal cognitive heuristics are those that are good at producing true beliefs.
> (4) Biological evolution favors those mechanisms that are good at producing true beliefs.

So far, I have argued that there is good reason to believe (1'). Consider now (2), the claim that if some trait is the result of natural selection, there is good reason to believe it is optimal. An objection to this premise is that even if a trait is the result of natural selection, it is not always the case that the trait was itself selected for, and, if a trait was not selected for, there is no reason to believe that it is optimal. Elliott Sober makes a helpful and important distinction between selection *of* and selection *for*.[26] A trait can be the *result* of natural selection without the effects of that trait having caused it to be selected. If this is the case, there has been selection *of* that trait without there being selection *for* it—"selection of does not imply selection for."[27] If, in contrast, a trait is selected because of its effects, then there is selection *for* this trait (there is, of course, selection *of* this trait as well—selection *for does* imply selection *of*). I call the process whereby there is selection *of* a trait without

[26] *Nature of Selection*, pp. 97-102; see especially the helpful picture on p. 99. See also *Adaptation and Natural Selection*, p. 9, for Williams' similar distinction between designating something as "the means or mechanisms for a certain goal or function or purpose" compared to designating something using "words appropriate to fortuitous relationships such as a cause and effect."

[27] *Nature of Selection*, p. 100.

there being selection *for* it free-riding.[28] Examples of free-riders include a spandrel (a trait that is an architectural by-product of set of traits that was selected for),[29] a "piggyback" trait (a trait that was selected because it is associated with some other trait that was selected for, for example, a trait that is the result of pleiotropy in the standard genetic sense), and an exaptation (a trait that emerges when a trait previously selected for is used to perform some new function).[30]

One might argue that the existence of these phenomena causes problems for the claim that evolution favors truth-tropic heuristics because it undermines the premise that natural selection will choose optimal traits (2). On the naive view of natural selection, that humans have a trait is reason to believe that it is optimal. But this is a mistake; if a trait is the result of one of the forementioned evolutionary phenomenon (that is, if the trait is a free-rider), then it was *not* selected for because it is optimal. Consider the example of the chin given by Gould and Lewontin. Why do we have a chin? On the naive adaptationist view, *because* we have a chin, having a chin must be selectively advantageous. But, according to Gould and Lewontin, the chin is simply the result of the selective advantage of two different jaw-related growth fields (the alveolar and mandibular growth fields). Having these jaw-related growth fields was selected *for* and having a chin of a certain type was simply an architectural side effect of them, that is, the chin is a spandrel. The chin, while unquestionably the result of evolution, was not selected for by natural selection.[31] The

[28] Sober, *ibid.*, p. 24, uses the term 'pleiotropy' for the process I call free-riding. Sober uses 'pleiotropy' to refer to situations in which a *cluster* of genes have more than one effect on the phenotype. One problem with this term is that its standard genetic sense denotes situations in which a *single* gene has multiple phenotypic effects. Sober's use of 'pleiotropy' for the broader notion could be derived by combining the standard genetic account of pleiotropy with Richard Dawkins's (non-standard) definition of a gene as a "portion of chromosomal material which potentially lasts enough generations to serve as a unit of selection"; see Dawkins, *The Selfish Gene* (Oxford: Oxford University Press, 1976), p. 30 and *passim.* Dawkins definition of 'gene' has been strongly criticized by Richard Lewontin, "Caricature of Darwinism," *Nature* 266 (March 17, 1977), pp. 283-284; and Gunther Stent, "You Can Take the Ethics Out of Altruism but You Can't Take the Altruism Out of Ethics," *Hastings Center Report* 7 (1977), pp. 33-36. I adopt the term 'free-riding' to avoid confusion due to the use of 'pleiotropy.'

[29] "Spandrels of San Marcos."

[30] Steven Jay Gould and E. S. Vrba, "Exaptation—a Missing Term in the Science of Form," *Paleobiology* 8 (1982), pp. 4-15.

[31] "Spandrels," p. 256 (page references from *Conceptual Issues in Evolutionary Biology*).

point about cognitive heuristics is that even if we can be sure that our cognitive mechanisms are the result of natural selection, we cannot be sure that they were selected *for* as a result of these traits themselves exhibiting selective advantage; after all, they might be free-riders. Lewontin, for example, suggests that all of our cognitive mechanisms are spandrels. He says that:

> there may have been no direct selection for cognitive ability at all. Human cognition may have developed as the purely epiphenomenal consequence of the major increase in brain size, which, in turn, may have been selected for quite other reasons.[32]

Gould makes a similar claim. He says that he does not doubt that:

> the brain's enlargement in human evolution had an adaptive basis mediated by selection. But I would be more than mild., surprised if many of the specific things it now can do are the product of direct selection "for" that particular behavior. Once you build a complex machine, it can perform so many unanticipated tasks.[33]

Note the difference between the argument against (2) that appeals to free-riders and the section III argument against (1) that appeals to non-selectionist evolutionary forces. The argument I considered and rejected in section IV is that natural selection is not the cause of our having the innate cognitive mechanisms that we have because other evolutionary forces besides natural selection might have been the cause of the evolution of cognitive heuristics. The argument I am currently considering grants that natural selection is the cause of our having the innate cognitive mechanisms that we do but challenges the claim that this means the mechanisms are optimal. This challenge turns on the existence of processes of natural selection that cause certain traits to be selected without their having been selected *for*.[34]

A defender of (2), the claim that natural selection favors the most optimal traits, might deny that any trait that is supposed to be the result of selection *of* but not selection *for* is a trait at all. Support for this claim might be drawn from Gould and Lewontin's critique of

[32] "Evolution of Cognition," p. 244.

[33] Steven Jay Gould, "Panselectionist Pitfalls in Parker and Gibson's Model of the Evolution of Intelligence," *Behavioral and Brain Sciences* 2 (1979), p. 386.

[34] Stich, *Fragmentation*, pp. 63-90, discusses variants of both objections—to (1) and (2)—but fails to be clear about the difference between non-selectionist evolutionary forces and selectionist evolutionary forces that might cause the selection of non-optimal traits.

the way that adaptationists atomize organisms into traits.[35] In the case of the chin, the defender of (2) could deny that the chin is a separate trait at all. In general, this move would deny the status of being a trait to anything that was not selected for. This would have the result that every trait would be selected for, thereby blocking the objection to the premise that natural selection selects optimal traits (2) based on the claim that some selected traits might not be optimal since they were not selected for. This response fails because only through a definitional fiat—that is, by simply not counting non-optimal cognitive heuristics as traits—does it save the claim that every trait is selected for. The important issue, however, is not what counts as a trait but whether any cognitive mechanisms are non-optimal. Changing a definition does nothing to settle this question.

Before evaluating the objection to (2) based on the distinction between selection of and selection for, I will broaden it. There are other ways besides being a free-rider that a trait could be selected without it being optimal. The trait might be a neutral trait (that is, it might have been selected not because of any features it has but because it had no negative features) [36] it might be the result of meiotic drive (a phenomenon that occurs when an allele "stacks the deck" in its own favor by making copies of itself in more than the usual fifty percent of the cells that are involved in meiosis, the process that makes gametes, for example, in humans, sperm and eggs),[37] or it might have been selected because of heterozygote superiority (a phenomenon whereby even if a homozygous trait is the most selectively advantageous, it might not be selected if it is a recessive allele and if its associated heterozygote trait is less adaptive than the homozygote pair of the dominant allele[38]).[39] Unlike the non-selectionist forces discussed earlier, these evolutionary

[35] "Spandrels," p. 256.

[36] See Mooto Kimura, "The Neutral Theory of Evolution," *Scientific American* 240:5 (1979), pp. 98-126.

[37] J. Crow, "Genes That Violate Mendel's Rules," *Scientific American* 240:2 (1979), pp. 134-146.

[38] These technical terms need some explaining. Genetic factors come in sets of traits that can be exhibited in organisms. These sets of traits are called *alleles*. For example, the alleles for human eye color include a blue allele, a green allele, etc. Only two alleles for the same feature (for example, eye color) can be present in a single organism—one allele from the person's father and one from the person's mother. If an organism has the same allele from both parents, it is said to be a *homozygote*; if an organism has a different allele from each parent, it is said to be a *heterozygote*. Some alleles are dominant and others are

phenomena (including free-riding) have very much to do with natural selection. Natural selection can explain why we have, for example, a trait that results from its having "ridden piggyback" on some other trait but this explanation does not invoke selection *for* this trait; rather, natural selection explains selection *for* some other trait that in turn explains the selection *of* the piggyback trait.

With all of these various evolutionary phenomena that undermine a simple picture of natural selection, (2), the claim that evolution selects for optimal traits is in trouble. A possible way of saving this claim—a move that parallels my answer to the above objection to (1) by modifying it to get (1')—would be to admit that natural selection sometimes, due to the evolutionary phenomena discussed above, fails to select optimal traits, but deny that these phenomena play a role in the evolution of functionally complex structures such as human cognitive mechanisms. This would save the claim that cognitive mechanisms resulting from natural selection are optimal. Thus (2) could be replaced with:

> (2') With respect to functionally complex structures (such as human cognitive mechanisms), natural selection favors optimal traits.

But what reason is there for thinking that such evolutionary phenomena as pleiotropy, heterozygote superiority, meiotic drive, etc. (phenomena that, unlike, for example, genetic drift, are part of natural selection) do not operate in the evolution of functionally complex structures? A plausible answer parallels the one given above to the objection that non-selectionist evolutionary forces sometimes explain the evolution of cognitive mechanisms. In the previous section, I argued that natural selection is the only explanation for the evolution of functionally complex structures. The parallel response to the challenge that traits resulting from natural selection are not always selected for (and thus not necessarily optimal) is that the evolution of functionally complex structures can only be explained by

recessive. If an organism is a heterozygote, the dominant allele will be expressed (for example, if brown eyes are dominant and the person is a brown eye/blue eye heterozygote, then she will have brown eyes) while the recessive allele will not be.

[39] See Philip Kitcher, *Vaulting Ambition: Sociobiology and the Quest for Human Nature* (Cambridge: MIT Press, 1985), p. 215, and A. Templeton, "Adaptation and the Integration of Evolutionary Forces," in *Perspectives on Evolution*, R. Milkman, ed. (Sunderland, MA: Sinauer, 1982).

these traits having been selected *for* because only natural selection can explain the sort of functional complexity that they exhibit.

Note that this response does not require that there has been selection *for* the properties associated with every *possible* way of describing every part of the reasoning faculty. There are an infinite number of ways to describe an organ or its parts that do not correspond to a function it performs. Dr. Pangloss's example of the nose being made to hold eyeglasses[40] is such an example as are Pinker and Bloom's examples of the redness of blood and the fact that humans have a prime number of digits of each limb.[41] Clearly, such non-functional "traits" are not selected for. The reasoning organ and its functional parts are not, however, non-functional. Rather, a highly specialized organ the reasoning faculty must be the result of selection for its functions. The specificity with which the reasoning faculty matches specific cognitive tasks that are important for humans to perform makes it highly unlikely (*contra* Gould and Lewontin) that this faculty is simply the result of something like an increase in brain size.

While this response works against the Gould-Lewontin picture of the evolution of the human reasoning faculty, it does not work against a more specific account of how a cognitive mechanism could have been selected without having been selected *for*. Suppose that a cognitive heuristic is selected for because it performs a particular function in certain circumstances. The same heuristic might well perform a function that was not selected for in other circumstances. This function, not having been selected for, might well be non-optimal. To make this concrete, consider the Linda example discussed in Chapter One.[42] Imagine that the heuristic for making *plausibility* judgments (such as whether it is more *plausible* that Linda, a former philosophy major, is now a bank teller or both a feminist *and* a bank teller) is selected for because of its use for making plausibility judgments. This

[40] The example is cited in "Spandrels," p. 254.

[41] "Natural Language and Natural Selection," p. 710.

[42] See Amos Tversky and Daniel Kahneman, "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," *Psychological Review* 90 (October 1983), pp. 293-315, for discussion of probability judgments.

heuristic might also be called into use for making *probability* judgments (whether it is more *probable* that Linda is a bank teller or both a feminist and a bank teller). Having not been selected for to make probability judgments, the heuristic may not be the optimal heuristic for making such judgments. In fact, the plausibility heuristic is not the optimal heuristic for making probability judgments. While it might be more *plausible* that a philosophy major would be a feminist bank teller than a bank teller (whether or not she is a feminist), it cannot be more *probable* that anyone (regardless of her major) would be a feminist bank teller than a bank teller, because all feminist bank tellers are necessarily bank tellers. The general point is that a heuristic that was selected but *not* selected *for* can be non-optimal; the selection forces behind this fact undermine the simple picture of natural selection and count against (2'). While selection for is the only plausible explanation of the evolution of functionally complex structures, that such a structure has been selected for does not guarantee that every feature of it will have been selected for.

Looking back to Pinker and Bloom's argument concerning the evolution of the language faculty, it is interesting to note a difference between their argument and the parallel argument about the evolution of the reasoning faculty. While Pinker and Bloom seem to implicitly argue that the language faculty was selected for, they are not troubled by the fact that various functional features of the language organ were *not* selected for; in fact, they admit that certain features of the language organ were *not* selected for.[43] However, when the Williams-Pinker-Bloom argument is applied to the evolution of the reasoning faculty, such an admission becomes problematic. Admitting that a feature of the reasoning faculty (for example, a cognitive heuristic) was selected but not selected for, opens up the possibility that this feature is not optimal and thus could be the source of irrationality. This difference between the language organ and the reasoning organ—namely, that non-optimality is not an interesting problem with respect to language but it is with respect to

[43] "Natural Lan' age and Natural Selection," pp. 715-720.

reasoning—is explained by the fact that the norms of language are indexed to actual linguistic competence while the norms of reasoning are not.[44]

Natural selection is the only explanation for the evolution of functionally complex structures like our reasoning faculty. Further, selection for such complex structures is the only plausible story behind this natural selection process—Gould and Lewontin's suggestion that our cognitive mechanisms might result from selection for an increase in brain size is mistaken. That selection for is the only plausible story behind complex structures does not, however, mean that every feature of our cognitive mechanisms was selected for; such features may well result from selection for other features of the reasoning faculty, and as such, these features may be not be optimal.

VI.

Nothing I said in the previous section proves that in fact those cognitive mechanisms produced by natural selection are not optimal. Suppose, for the sake of argument, that all of our cognitive mechanisms have been selected for and that all of the traits of such mechanisms were selected for because of the functions they perform. Even granting this, there remains the further question of whether this entails that human cognitive heuristics are optimal, more particularly, whether they will be optimal in the sense that they will produce true beliefs (3). I will discuss this question in terms of the metaphor of a filter. The question is whether natural selection[45] provides a filter of the appropriate grain size to select heuristics that produce true beliefs.[46] The question can be seen as a challenge to either (2) or (3): if directed at (2), the challenge is that natural selection, even setting aside

[44] See my discussion of this difference in Chapter Two, section IV.

[45] For the remainder of this chapter, unless otherwise noted, by 'natural selection,' I mean natural selection "at its best," that is, when it is selecting *for* traits. This is a legitimate shorthand to adopt for this section because I am assuming, for the sake of argument, that natural selection has selected for every feature of our cognitive mechanisms.

[46] For more on the filter as a metaphor for natural selection, see Peter Lipton and Nicholas Thompson, "Comparative Psychology and the Recursive Structure of Filter Explanations," *International Journal of Comparative Psychology* 1 (Summer 1988), pp. 215-244.

the objections of the previous section, will not be able to select the optimal heuristics; if directed at (3), the challenge is that truth-tropic heuristics are not optimal. For simplicity, I will not distinguish between these different ways of describing the challenge; rather, I will characterize it as directed at the conjunction of (2) and (3), namely, the claim that natural selection favors heuristics that yield true beliefs. I will consider two objections that go together—the first says that the filter of natural selection is too coarse-grained (that is, it selects some heuristics that mostly produce false beliefs) and the second says that it is too fine-grained (that is, it selects some heuristics that produce very few true beliefs).[47]

Even at its best, biological natural selection, for all its power, is not omnipotent; rather it is pragmatic, that is, result oriented. To simplify the picture, when faced with two mutually exclusive alternatives in a particular environment, natural selection will choose the alternative that will do better in that environment. Being better may amount to being faster or more precise, or some combination thereof, depending on the situation. How do these facts carry over to the case of cognitive heuristics? If natural selection favors those heuristics that produce true beliefs, then a heuristic (call it H_T) that produces true beliefs ought to be selected for over a heuristic (call it H_F) that does not. But since natural selection selects on the basis of effects, if H_T and H_F have the *same* effects, natural selection will not be able to choose between them, that is, either *neither* will be selected for (and natural selection will prove too fine-grained to produce truth-tropic heuristics), or *both* will be selected for (and natural selection will not be fine-grained enough), or natural selection will choose between them at random. Biological evolution cannot simply select for those heuristics that are truth-tropic since whether a heuristic is truth-tropic is underdetermined by the heuristic's effect in a particular environment.[48]

[47] These objections can be seen roughly as abstracted versions of the three specific problems Sober discusses as to why natural selection involves too coarse-grained (my term, not his) a filter to select for the scientific method; see "Evolution and Rationality," p. 98. See also *Fragmentation*, pp. 60-63; and "Could Man Be an Irrational Animal?," pp. 257-258.

[48] It is even underdetermined by the heuristic's global effect. It follows from the fact that natural selection is unable to distinguish between two heuristics with the same effects in *any* given environment that natural selection is unable to distinguish between two heuristics with the same effects in *every* possible

Sober gives a nice illustration of the problem that the pragmatic nature of natural selection causes for the conjunction of (2) and (3), the claim that natural selection favors heuristics which yield true beliefs. He presents three heuristics:

> Induction: "If **n/m** of the observed As have been **B**, infer that **n/m** of the remaining As are Bs."[49]
>
> Counterinduction: "If **n/m** of the observed As. . . [have been] **B**, infer that 1-(**n/m**) of the remaining As are Bs."[50]
>
> Mixed Strategy: "[Use induction] . . . if you are making an inference before the year 2000; use [counterinduction] . . . if you are making an inference after the year 2000."[51]

Note that, before the year 2000, natural selection cannot, on the basis of the beliefs these heuristics would yield, differentiate between induction and the mixed strategy. Suppose that, in fact, induction is more truth-tropic than the mixed strategy. Since these heuristics have the same effects before the year 2000, natural selection is unable to choose between them on the basis of the extent to which they are truth-tropic. The point is that the filter of natural selection is too coarse-grained to allow just truth-tropic heuristics to pass through; other heuristics that have desirable effects will be in the precipitate as well.

Sober has a response to this criticism. He notes that natural selection selects for more than just external effects (that is, what the trait does with respect to the environment); it also selects for internal effects (that is, how the trait operates from the *processing* point of view).[52] Natural selection can be thought of as drawing on considerations similar to those that humans use, say, when buying a computer: besides comparing what functions one computer can perform compared to another, we also consider the computers' size, speed, cost, etc. These features are analogous to the internal effects of a trait that natural selection

environment; see "Evolution of Rationality," p. 102. Sober compares induction (as described below) to the adding strategy (also below). From the point of view of their effects, natural selection cannot, the objection goes, distinguish between induction and the adding strategy because they have the same effects in every possible environment.

[49] "Evolution and Rationality," p. 101.

[50] *Ibid.*

[51] *Ibid.* Sober here draws from Nelson Goodman's well-known "grue" paradox; see Goodman, *Fact, Fiction and Forecast*, fourth edition (Cambridge: Harvard University Press, 1983) pp. 59-83.

[52] The idea of focusing on the processing point of view when looking at natural selection comes from *Adaptation and Natural Selection*, *passim*, especially p. 33.

considers. While induction and the mixed strategy have the same *external* effects, they have different *internal* effects—the mixed strategy requires the use of a clock of some sort to tell whether the year 2000 has been reached, while induction does not. Induction also has the same external effects as the following heuristic:

> Adding strategy: If **n/m** of the observed As have been **B**, add up the first 10^{10} integers. From this total, then, subtract the first 10^{10} integers. Take the resulting number and add it to **n/m** and infer that this number represents the proportion of the remaining As which are **B**.[53]

While the adding strategy and induction have the same external effects, (both before *and* after the year 2000) their internal effects are quite different; the adding strategy requires a large amount of processing time and a sophisticated adding machine while induction requires neither. Given the relative internal effects, induction is preferable to the adding strategy and the mixed strategy since it is more efficient and better focused on the task it is supposed to accomplish.

How is including consideration of internal effects supposed to save the claim that natural selection favors truth-tropic heuristics from the objection that the filter of natural selection is too coarse-grained to select for truth-tropic heuristics? The idea is that natural selection is more fine-grained than critics of the conjunction of (2) and (3) would imagine. Besides looking at external effects, natural selection looks at internal effects such as speed, memory used, efficiency, etc.; in general, internal economy is a virtue, a virtue that natural selection will recognize. Internal economy is not, however, preferable if it involves the sacrifice of a great deal of external performance. In this sense, natural selection is also like a person on the market for a computer: all else being equal, a person would choose the smaller, cheaper, faster computer, but would not opt for such internal economy if doing so required that the computer would be unable to, say, function as a word processor.

That natural selection is fine-grained enough so that internal factors can play a role in selection does not, however, show that natural selection is fine-grained enough to select for

[53] "Evolution and Rationality," p. 102.

truth-tropic heuristics. In fact, internal effects are often in tension with truth-tropicity; natural selection will often prefer the quick-and-dirty approach to solving a problem posed by the environment rather than the truth-tropic one. While internal effects are no doubt involved in natural selection, it is wrong to think that consideration of these effects guarantees, or even makes it more likely, that truth-tropic heuristics will be selected for and non-truth-tropic heuristics will not be.

I will now consider a ubiquitous example that is supposed to show that natural selection might select for a non-truth-tropic heuristic over a truth-tropic one.[54] In an experiment performed by John Garcia, et al., rats were shown to develop a strong aversion to food of a distinct flavor that they have been fed before being subject to substantial (that is, sickness causing) doses of radiation.[55] Later, Martin Seligman discovered (the hard way) a similar effect in humans that he called the Sauce Béarnaise effect. Six hours after eating filet mignon with Sauce Béarnaise, Seligman got sick due to a stomach virus; subsequently, he was unable to eat the sauce (although he had no problem eating filet mignon, eating in the same environment, eating off of the same plates, etc.).[56] One natural way to describe this phenomenon is to say that humans and rats seem to be following a heuristic such as:

> Cautious heuristic: If I eat food with a distinct taste and subsequently get sick, infer that food which tastes like this will make me sick.[57]

Although the Sauce Béarnaise did not in fact *cause* Seligman to get sick, the cautious heuristic would explain why he subsequently became unable to eat Sauce Béarnaise—he inferred, from the mere conjunction of the sauce and his sickness, that the sauce *caused* the sickness. This is a bit of a simplification. The Garcia effect and the Sauce Béarnaise effect

[54] The example is discussed for similar purposes in *Fragmentation*, pp. 61-63, and "Could Man Be an Irrational Animal?," pp. 256-260, among other places.

[55] J. Garcia, B. K. McGowan, and K. F. Green, "Biological Constraints on Conditioning," in *Classical Conditioning: Current Research and Theory*, volume 2, A. H. Black and W. F. Prokasy, eds. (NY: Appleton-Century Crofts, 1972), pp. 3-27.

[56] Martin Seligman and Joane Hager, *The Biological Boundaries of Learning* (NY: Appleton-Century Crofts, 1972), p. 8.

[57] Stich,"Could Man Be an Irrational Animal?," pp. 257-258, seems to imply that such a heuristic is being followed.

seem to be non-cognitive heuristics akin to reflexes. We have no reason to think that Seligman consciously inferred that the Sauce Béarnaise caused his sickness in the same way that I often consciously infer q from p and if p, then q; in fact, we have no reason to believe that any conscious cognitive computations were involved in the development of Seligman's aversion to the sauce. For my purposes, however, I want to discuss a (perhaps imaginary) effect like the Garcia effect and the Sauce Béarnaise effect except that the imaginary effect is a conscious cognitive one. All of these effects—the Garcia effect, the Sauce Béarnaise effect and the imaginary effect—can be characterized as fostering *overdetection*; heuristics of this form are more likely to mistakenly identify non-poisonous foods as poisonous (make "false positives") than to fail to identify as poisonous foods that actually are (make "false negatives").[58] Henceforth, I will call the resulting behavior of the imagined cognitive heuristic the Garcia effect since it has the same general form as the Garcia effect in rats. The important point is that natural selection might select a heuristic that gives rise to the Garcia effect and, as I will discuss below, such a heuristic, contrary to the conjunction of (2) and (3), would not be not truth-tropic.

Returning to the imaginary cognitive version of the Garcia effect, let us suppose that greater than fifty percent of the sicknesses that affect humans have nothing to do with food, distinctly flavored or otherwise. If this is true, then the cautious heuristic will tend to produce more false beliefs than true beliefs. This was the case with Seligman: the sauce did not cause him to be sick—it was the virus—but he was cautious and "blamed" the

[58] For other discussions of heuristics of this form, see Jerry Fodor, "Psychosemantics," in *Mind and Cognition: A Reader*, William Lycan, ed. (Oxford: Basil Blackwell, 1990), pp. 330-332. For a critique of it see Ned Block, "Advertisement for a Semantics for Psychology," in *Midwest Studies in Philosophy*, volume 10, P.A. French, et al., eds., (Minneapolis: University of Minnesota Press, 1986), note 65, pp. 673-674. Fodor is concerned with giving an evolutionary account of *meaning*, not an evolutionary argument for rationality (an account he subsequently rejects—see, for example, "A Theory of Content" in *A Theory of Content and Other Essays* (Cambridge: MIT Press, 1990), pp. 51-87), but some of the points he makes are relevant here (see below). For a more sophisticated version of an evolutionary account of meaning, see Ruth Millikan, *Language, Thought and Other Biological Categories* (Cambridge: MIT Press, 1987); Papineau, *Reality and Representation*; and Daniel Dennett, *The Intentional Stance*. There is not necessarily a connection between the truth of an evolutionary theory of meaning and the success of the evolutionary argument for the rationality thesis.

sauce. The cautious heuristic is thus not truth-tropic. The crucial point is that despite all this, the cautious heuristic could still be a perfectly good heuristic for natural selection to choose; it might be more adaptive (that is, it might do a better job of fostering reproductive success) in some environments for an organism to be cautious when it comes to the threat of food poisoning. Natural selection will thus sometimes fail to produce truth-tropic heuristics; the conjunction of (2) and (3) is thus false.

There are two general problems with interpreting cognitive mechanisms that arise here. First, it is in general difficult to characterize a cognitive mechanism on the basis of the behaviors it causes since behavior underdetermines the underlying mechanism. If Seligman avoids eating Sauce Béarnaise due to some cognitive process, he might be following the cautious heuristic. He might, however, be following one or the other of the following heuristics:

> Probabilistic heuristic: If I eat food that has a distinctive taste and subsequently get sick, I should believe that food which has this taste is more likely to make me sick than most food I encounter.[59]
>
> Two-conditional heuristic: If I eat food that has a distinctive taste and subsequently get sick, I should believe that food with this taste will make me sick; if I do not get sick, I should believe that food with this taste will *not* make me sick.[60]

Both of these heuristics would explain Seligman's behavior as well as does the cautious heuristic.

The second problem is that even if neuroscientists discover the actual cognitive mechanism behind the Garcia effect, there still is a difficulty in characterizing it. This problem is often discussed in terms of what the frog's eye tells the frog's brain because the underlying mechanism in frogs is fairly well understood.[61] Even if neuroscientists could

[59] This is modification of what Feldman, "Rationality, Reliability, and Natural Selection," p. 221, calls strategy A. He says that I should believe that this taste *might* make me sick, but this seems trivially true since any food *might* make me sick. See also "Psychosemantics," p. 330, for Fodor's version of this strategy applied to another context: "[T]hat just might be a predator and I'm taking no chances."

[60] This is a paraphrase of what Feldman, "Rationality, Reliability, and Natural Selection," p. 222, calls strategy B.

[61] J. Y. Lettvin, et al., "What the Frog's Eye Tells the Frog's Brain," *Proceedings of the Institute of Radio Engineers* (1959), pp. 1940-1951.

come up with a complete description of this mechanism, there would still be the question of whether to interpret the mechanism as a fly detector or whether to call it an object-of-size-X-travelling-at-speed-y detector. The principle of charity would suggest that the latter characterization is the most appropriate for the mechanism since, on this characterization, the mechanism never fails to work.[62] A teleological account would, however, suggest that the former account is the most appropriate characteiization for the mechanism since the mechanism was selected and retained because it is good at detecting flies (not because it is good at detecting certain-sized pellets).[63] Along these lines, the cautious heuristic could either be interpreted as a taste-that-once-preceded-sickness detector or, following the teleological account, as a poison detector.

Both of these problems (that the mechanism is underdetermined by its behavior and that even if the mechanism is isolated, there is no obviously correct characterization of it) interact with whether the Garcia effect is properly explained by the cautious heuristic and is, hence, an example of a non-truth-tropic heuristic that has been selected by natural selection. A friend of the conjunction of (2) and (3) could defend it against the above problems by criticizing the cautious-heuristic analysis of the Garcia effect. This could be done by noting that the probabilistic and the two-conditional heuristics would explain the rats' behavior as well as does the cautious heuristic but that these two other heuristics *are* truth-tropic.[64] The cautious heuristic is not truth-tropic because false beliefs are produced in most of the cases in which it is used to generate beliefs. The probabilistic heuristic will be used in the same instances as the cautious heuristic (that is, when distinctly flavored food is followed by sickness), but the probabilistic case will yield beliefs that are true, that

[62] See W. V. O. Quine, *Word and Object* (Cambridge: MIT Press, 1960) for the classic formulation of the principle of charity. See Davidson, *Inquiries*, and Dennett, *Intentional Stance*, for attempts to apply the principle of charity to principle of reasoning. For criticisms of such attempts, see *Fragmentation*, and "Rationality and the Limits of Cognitive Science," Chapter Four.

[63] Among the many who have gotten philosophical mileage from this frog-fly example are Daniel Dennett, "When Frogs (and Others) Make Mistakes," in *Intentional Stance*, pp. 103-116; and Ruth Millikan, "Biosemantics," *Journal of Philosophy* 86 (1989), pp. 281-297.

[64] This is roughly the strategy taken by Feldman in "Rationality, Reliability, and Natural Selection."

is, distinctly flavored food consumed before getting sick is more likely to be a source of sickness than food whose consumption does not precede sickness. The two-conditional heuristic will be used far more frequently than the cautious heuristic (since it will be used whenever any distinctive tasting food is eaten), but it will generate true beliefs most of the time since it will usually yield the inference that a distinctively tasting food that does not cause sickness after one consumption does not cause sickness in general.

The argument against the conjunction of (2) and (3) is that sometimes natural selection rejects truth-tropic heuristics in favor of non-truth-tropic ones. It proceeds by example: the cautious heuristic is not truth-tropic yet it gets selected. The response under consideration is that there is no reason to think the cautious heuristic is used rather than the probabilistic heuristic or the two-conditional heuristic. Since the latter two are truth-tropic, there is no reason to think humans use a heuristic that is not truth-tropic. Thus the opponent of the conjunction of (2) and (3) has failed to produce an adequate counterexample to the claim that natural selection produces truth-tropic heuristics.

A defender of (2) and (3) might, however, reply that this response misses the general point. The particulars of the example of the Garcia effect are not important. Even if it turns out that the two-conditional heuristic is what causes the Garcia effect, natural selection *could* have chosen a heuristic like the cautious one over some more truth-topic heuristic *if* the cautious heuristic was selectively advantageous, that is, if it led to greater reproductive success. More generally, as Stich puts it, "Natural selection *could* perfectly well lead to inferential strategies which generally get the wrong answer, but are right when it counts most, just as it leads to aversions to foods most of which are harmless and nourishing."[65] The point is that natural selection selects for heuristics that are adaptive and it is certainly possible for heuristics that are not truth-tropic to be the most adaptive; whether or not the cautious heuristic is actually what is behind the Garcia effect, it remains a possibility that a non-truth-tropic heuristic might be preferred over a truth-tropic one.

[65] "Could Man Be an Irrational Animal?," p. 258, emphasis added.

What might a defender of the argument for (4), the claim that biological evolution favors truth-tropic heuristics, say in response to this objection? I see three possible responses. The first is to try to generalize from the above response to the cautious-heuristic case. The general response would be that, when presented with a heuristic that is claimed to be adaptive but not truth-tropic, try to argue that the heuristic at hand has been incorrectly characterized and that, correctly characterized, the heuristic really is truth-tropic.[66] Call this the reinterpretation strategy. The second response is to argue that whatever the mechanism actually is, it can be construed as truth-tropic (for example, even if the cautious heuristic is the right mechanism, it can be construed as a taste-that-once-preceded-sickness detector rather than a poison detector). The third response to the criticism of the defense of the conjunction of (2) and (3) is to weaken (3) so that it does not claim that all optimal heuristics produce true beliefs. I will consider each of these responses in turn.

Consider first the reinterpretation strategy. Suppose, without loss of generality, that heuristics are of the form:

If p occurs/is believed/etc., then do/believe/infer/etc. q.

Reinterpretation can be performed, for example, by weakening the consequent of the conditional (as with the probabilistic heuristic) or by coupling the conditional with another heuristic (as in the two-conditional heuristic). There are other possible ways of reinterpreting heuristics that initially seem not to be truth-tropic so that they seem truth-tropic. There is, however, a problem with this strategy. Reinterpretation may not always be motivated—why should we think that truth-tropic heuristics are more likely to be selected than those that are not truth-tropic? What possible reason do we have for believing that humans (or rats) use the probabilistic or two-conditional heuristic rather than the cautious one? Besides the principle of charity that I argue against in Chapter Four, the only general motivation I can see for believing that truth-tropic heuristics are more likely to be

[66] This is, more or less, Fodor's strategy in "Psychosemantics," p. 331: "so ascribe beliefs that, as the organism's evidence that P becomes arbitrarily transparent, the likelihood that the organism believes that P approaches arbitrarily close to unity. . . ."

possessed is if one already believes that truth-tropic heuristics are more likely to be selected by biological evolution. But this is precisely what is at issue here.[67] Further, whatever reinterpretation is offered, it remains a possibility that a heuristic will be more selectively advantageous than another that is more truth-tropic than it is.

To review, I am considering the claim that natural selection favors truth-tropic heuristics. I am examining the Garcia effect as a possible counterexample to this claim as well as the attempt to defuse the example by denying that the Garcia effect constitutes a counterexample at all. Maybe the defusing attempt works in the particular example, but the general difficulty remains: sometimes the most adaptive heuristic might not be truth-tropic. The first answer to the general difficulty I considered was that the sort of reinterpretation of non-truth-tropic heuristics as truth-tropic heuristics is always possible and plausible. But, there are two problems with this answer. First, it just begs the question since there is no reason for believing such a reinterpretation is justified in general unless you already believe that natural selection favors truth-tropic heuristics. Second, all that matters for the objection to go through is the *possibility* that a less truth-tropic heuristic will be selected over a more truth-tropic one; the reinterpretation strategy has nothing to say to this problem.

The second response to the general difficulty has the virtue of addressing the second problem with the reinterpretation strategy. Recall that the second response is that whatever the cognitive mechanism is, it can be interpreted as truth-tropic. Just as the mechanism that sends messages from the frog's eye to the frog's brain is infallible if characterized as an object-of-size-X-travelling-at-speed-y detector (rather than as a fly detector), so too the cautious heuristic (assuming it is the mechanism behind the Garcia effect) is truth-tropic if it is characterized as a taste-that-once-preceded-sickness detector (rather than a poison detector). This reply, in effect, challenges the claim that it is possible for a non-truth-tropic

[67] This is similar to an objection raised by Block, "Advertisement," pp. 673-674, note 65, against Fodor.

heuristic to be selected over a truth-tropic one by claiming that seemingly non truth-tropic mechanisms can be construed as truth-tropic.

Suppose that the cautious heuristic is in fact truth-tropic, (namely, that it is a taste-that-once-preceded-sickness detector). This does not mean that in general natural selection selects for truth-tropic heuristics. Neither does the fact (if it is a fact) that any mechanism can be interpreted as truth-tropic show that natural selection selects for truth-tropic heuristics. What needs to be defended is the claim that mechanisms or heuristics resulting from natural selection *must* be interpreted as truth-tropic. Only by establishing this stronger claim can the conjunction of (2') and (3) be supported. I know of no successful way of defending this stronger claim.

The third attempt to support the argument for (4), the claim that biological evolution favors truth-tropic heuristics, retreats from the claim that natural selection selects for truth-tropic heuristics. The defender of (4) might admit that sometimes natural selection selects non-truth-tropic heuristics and sometimes it fails to select truth-tropic heuristics but deny that this happens very often. This could be done by modifying (3), the claim that optimal cognitive heuristics are truth-tropic as follows:

> (3') A very high percentage of optimal cognitive heuristics are truth-tropic.

This premise, together with the claim that natural selection favors the most optimal cognitive mechanisms (2'), entails the conclusion that a high percentage of the heuristics produced by natural selection will be truth-tropic. If this conclusion is right, then the argument for (4) might be salvageable.

This change to the argument for (4), that claim that evolution favors truth-tropic heuristics, specifically addresses the objection discussed above. Above, I made the claim that the defender of the conjunction of (2') and (3) needs to show that the heuristics resulting from natural selection *must* be interpreted as truth-tropic. Replacing (3) with (3') makes it possible to avoid this task—to support (3'), one need not establish that all heuristics possibly produced by natural selection are truth-tropic, only that most of them are. This is perhaps a more manageable task, but it is one that I will not explore since the argument for (4), even so modified, faces another serious objection. I turn to this objection in the next section.

VII.

Natural selection, even when it selects *for* a specific trait, can only choose between available alternatives; the best alternatives are, however, not always among the available ones.[68] A trait is available to an organism if it is physically possible for an organism of that species to possess it. Consider two examples: locomotion and neuronal communication. The most efficient, most adaptive device for getting around certain terrains mammals inhabit is the wheel, yet no mammal has evolved a wheel-like biological component. Having a nervous system that uses optical fibers would be faster, more efficient, and so on, than the nervous system of any known organism, yet no organism has evolved such a super-fast nervous system. Why have these better (perhaps the best) alternatives not evolved? The simple answer is that (so far as we know) no mutations have ever caused an organism to sprout wheels or synthesize optical fibers, therefore the wheel has never been able to compete against legs for selection as a form of locomotion and optical fibers have never been able to compete against axons and dendrites as a form of neuronal communication; neither was available as a candidate for selection.

The longer answer to why "better" alternatives like a nervous system made up of optical fibers have not evolved involves the concept of preadaptation and requires a bit of a digression. A general difficulty facing biological natural selection is the problem of complex organs. As I noted above, it is virtually impossible for a single mutation to occur that would produce a complex organ such as a bird's wing, because many different genes are involved. The probability that the necessary number of mutations would occur

[68] This objection is discussed in a somewhat different form in "Evolution and Rationality," pp. 110-111; and "Rationality, Reliability, and Natural Selection," p. 220.

simultaneously, in a coordinated fashion, is infinitely small. So it appears that natural selection cannot explain the evolutionary developnient of these organs.

This anomaly can be resolved by what biologists call "preadaptation."[69] As part of providing an account of how complex structures like wings develop, biologists often posit simpler intermediate structures that occurred before the wing. This does not resolve the anomaly unless these intermediate structures themselves have some adaptive advantage. If the intermediate structure had no adaptive function, then it would generally not be retained and so could not lead to a wing. Obviously, the half-wing could not enable its possessor to fly (otherwise, it would be a wing). There must be an adaptive advantage of a pre-winged bird having a half-wing structure. The biologist's answer is to give an account of the half-wing's preadaptive function. For e·ample, the half-wing may have been used for trapping insects. For biologists, the move to preadaptation allows them to explain how a complex organ may have evolved; it does so by claiming that the structure evolved from an "ancestor" of the organ that had a different function.[70]

Preadaptations are examples of how the evolution of traits does not occur out of nothing; it occurs "on top of" previously existing traits. Given the available preexisting traits (that is, the possible preadaptations), it is not surprising that neither a nervous system made of optical fiber nor a mammalian wheel ever evolved; the necessary preexisting structures from which a wheel could be produced by mutation never existed. The wheel and the optical fiber examples show how natural selection is only able to choose between available alternatives and how the best traits are often biologically unavailable. Even

[69] Gould and Vrba, "Exaptation," have argued for the use of 'exaptation' in place of 'preadaptation,' since the latter may seem to suggest that the resulting structure (that is, the preadaptation/exaptation) was in some way anticipated. Their point is well taken—that is, the connotation of foresight needs to be avoided—but his suggested term need not be adopted in place of the traditional term. It is enough to just underscore the following point: when I use the term 'preadaptation,' I do not mean to suggest any foresight on the part of natural selection.

[70] Of course, not every complex organ needs to be explained by preadaptation. Every (retained) mutation occurring along the way in the development of *some* complex organs ied to improvement in the *same* function. A possible example of this is the eye: each (retained) mutation along the way in the evolutionary development of the eye presumably led to improved vision. For discussion of the evolution of the eye, see *The Blind Watchmaker*.

though the wheel may be the best possible organ for mammalian locomotion in certain environments, it cannot be selected unless it is available and it will not be available if the necessary preexisting heuristics are not available. In fact, often the best heuristics (in terms of rationality) are not available because they cannot possibly be implemented on a system the size of the human brain and executed in a time as short as a human life.[71] The point is that those heuristics that always or usually yield true beliefs are less likely to be suitable for implementation in humans.[72]

To return to the argument for the claim that biological evolution selects for truth-tropic heuristics, how is the fact that natural selection only chooses between available heuristics relevant to this argument? This fact, presented in the form of what I call the availability objection, is supposed to count against the claim that a high percentage of the heuristics produced by natural selection will be truth-tropic (the conjunction of (2') and (3')). If truth-tropic heuristics are not, in general, available, then natural selection will not be able to produce a significant percentage of them.

A defender of the claim that biological evolution produces truth-tropic heuristics (4) might try to come up with some argument that truth-tropic heuristics *are* typically available for selection. I do not, however, know of any such argument that will work. A more promising approach, I think, is to try to modify (4) and the argument for it in order to take the availability objection into account. At this point, the argument for (4) is as follows:

> (1') The biological evolution of functionally complex structures (including human cognitive mechanisms) is caused by natural selection.
> (2') With respect to functionally complex structures (such as human cognitive mechanisms), natural selection favors optimal traits.
> (3') A very high percentage of optimal cognitive heuristics are truth-tropic.
> (4) Biological evolution favors those mechanisms which produce true beliefs.

[71] See Christopher Cherniak, "Undebuggability and Cognitive Science," *Communications of the Association for Computing Machinery* 31 (1988), pp. 402-412; and "Rationality and the Limits of Cognitive Science."

[72] Daniel Dennett, "Cognitive Wheels: The Frame Problem in AI," in *Minds, Machines and Evolution*, Christopher Hockaway, ed., (Cambridge: Cambridge University Press, 1984), pp. 146-147, calls such heuristics "cognitive wheels." By this term, he means a design proposal that, though very elegant and powerful, is highly unlikely to be instantiated in a biological system like the human mind. Heuristics that are ideally rational but not implementable on human brains can be thought of as cognitive wheels.

In order to save this argument in the face of the availability objection, (2') might be modified in the following fashion:

> (2") With respect to functionally complex structures (such as human cognitive mechanisms), natural selection favors the most optimal *of the available* cognitive mechanisms.

But now (3') needs to be modified as well:

> (3") A very high percentage of the most optimal *of the available* cognitive heuristics are truth-tropic.

The argument for (4) from (1'), (2") and (3") seems valid, but it is not sound. Compare (3') to (3"). The claim that optimal heuristics will be truth-tropic is much more plausible than that the most optimal *of the available* heuristics will be truth-tropic. The latter claim is contingent on there being truth-tropic heuristics in the set of available heuristics. What reason is there to think that this will be the case? A more plausible premise is:

> (3"') A very high percentage of the most optimal of the available cognitive heuristics will be more truth-tropic than the other available heuristics.

Together with (1') and (2"), this premise produces the following conclusion:

> (4') Biological evolution favors *the most truth-tropic of the available heuristics.*

This argument is (by design) immune from the availability objection. The question is whether this conclusion is an adequate replacement for the claim that biological evolution favors truth-tropic heuristics (4).

Recall that (4) is the first part the two-step evolutionary argument for (6), the claim that biological evolution produces rational heuristics. For this argument to be valid with (4') taking the place of (4), (5) will have to be modified as follows:

> (5') The most truth-tropic of the available heuristics is a rational heuristic.

Together (4') and (5') entail (6), the claim that natural selection favors rational heuristics. But are these premises true?

To begin, what exactly is it for a heuristic to be the most truth-tropic of those that are available? A tempting account is that a heuristic is the *most* truth-tropic of those available if it generates more true beliefs than any other available heuristic. This, however, cannot be right because, on this account, the following heuristic would be the mo:'t truth-tropic heuristic (or at least *as* truth-tropic as any other) of *any* set of available heuristics in which it was included:

All-inclusive strategy: If p is a possible belief, believe p.

No other heuristic could be more truth-tropic (on the account under consideration) than the all-inclusive strategy since no heuristic could possibly generate more true beliefs. But the all-inclusive strategy is clearly not a rational strategy since this strategy will sometimes sanction believing in p and not-p at the same time. If :he all-inclusive strategy is what is meant by the most truth-tropic, then (5') is false, since (5') says that the most truth-tropic of the available heuristics is rational. But if (5') is false, then the argument for (6) is not sound. The all-inclusive strategy must not be what is meant by the most truth-tropic.

A better account of what the most truth-tropic heuristic is would claim that the most truth-tropic heuristic is the one that can generate the best combination of the most true beliefs and the least false beliefs. But is there a general account of what such a best combination would consist in? Supposing such an account could be developed, would the claim that the most truth-tropic of the available heuristics is rational (5') be at all plausible? The Garcia effect is an example of why such a claim is unlikely to be true. Recall that in the case of the Garcia effect it seemed rational to follow the cautious heuristic (if I eat food with a distinct taste and subsequently get sick, infer that food which tastes like this will make me sick) or a behaviorally equivalent heuristic given that the result of a false negative (death by poisoning) is severe and the result of a false positive (aversion to a particular taste) is not, particularly given an environment where there is a great variety of food flavors available. The general point is that whether or not a heuristic is rational seems to be connected with something more than whether it generates true beliefs. Even if an account could be developed of what would be the best combination of accepting true beliefs and rejecting false ones, the claim that the most truth-tropic of the available heuristics is a rational heuristic (5') seems untenable since it deems rational the best of what might be a paltry set (that is, those that are available) of heuristics. Thus, the attempt to save (4), the claim that biological evoıution favors truth-tropic heuristics, by modifying it to be immune to the availability objection fails because the modified premise (4') is not strong enough to support (6), the conclusion that biological evolution produces rational heuristics.

VIII.

I began this chapter by spelling out a two-part evolutionary argument for human rationality. Up to this point, I have focused on the first part, the attempt to establish a connection between biological evolution and truth. The evolution of cognitive mechanisms is, I have argued, driven by biological natural selection, but, for at least three reasons, natural selection will not necessarily produce truth-tropic heuristics: biological natural selection does not guarantee that traits have been selected *for*, it can opt for some heuristics that are *not* truth-tropic and can *fail* to opt for some that are, and it can only choose from those heuristics that are available. These reasons suffice to break the connection between evolution and truth that the evolutionary argument tries to establish. There are two further modifications to this argument that I will turn to in this and the subsequent section. The first attempts to make the connection between evolution and rationality via reproductive success rather than truth; the second attempts to get to rationality via truth but to make the link with conceptual rather than biological evolution.

The evolutionary argument failed because natural selection cannot guarantee that the resulting cognitive mechanisms are truth-tropic. But there might be another way to deduce the rationality of cognitive mechanisms from the fact that they are the result of evolution, that is, there might be some way to make an evolutionary argument that does not involve truth. A plausible candidate is to make the link between evolution and rationalitv via reproductive success—although natural selection might not be a reliable producer of mechanisms that detect truth, it surely is a reliable producer of mechanisms that increase reproductive success. So, it seems perfectly reasonable that biological evolution will produce cognitive mechanisms that tend to increase reproductive fitness. Further, it seems rational to follow heuristics that increase reproductive fitness. From these two observations, it follows that evolution will tend to produce rational heuristics.

This argument seems a straightforward one when the effects of evolution on other realms are considered. Consider, for example, the visual system. Evolution favors visual mechanisms that lead to reproductive success. The only way to explain complex mechanisms like the human eye is to appeal to forces of natural selection that favor selective advantage in the form of increased reproductive success for the species.[73] Further, it is rational for an organism to behave in accordance with the mechanisms that increase reproductive success; it is clearly better and more rational to see with the visual system that natural selection picked for me than to try to see with some biological visual system that natural selection selected against. We seem therefore to have good reason to think that evolution will select the most rational visual system. The same seems true with respect to inferential systems.

This argument can be formalized as follows:

(RS4) Biological evolution favors heuristics that lead to the greatest reproductive success.
(RS5) A rational heuristic is one that leads to the greatest reproductive success.
(RS6) Therefore, biological evolution favors rational heuristics.
(RS7) Humans follow heuristics which are the result of biological evolution.
(RS8) Therefore, there is a strong reason to believe that humans are rational—in particular, there is a strong reason not to interpret psychological experiments as demonstrating human irrationality.

This argument shares the general (and the at least initially appealing) structure of the original evolutionary argument: it concludes that we have good reason to think humans are rational on the basis of the fact that the cognitive heuristics we use for reasoning are the result of biological evolution. At the same time, this modified argument avoids some of the problems of the original evolutionary argument, namely, it does not invoke truth or truth-

[73] See *Adaptation and Natural Selection* and *Blind Watchmaker* as well as my discussion above.

tropicity. The modified argument thus does not open itself to objections that the filter of evolution is too wide to precipitate only true beliefs or truth-tropic heuristics.

What, in particular, do we make of (RS4) and (RS5)? Given my arguments that many or most of our principles of reasoning are innate and constitute a complex mental organ, we have reason to believe that these principles are the result of natural selection. Since natural selection selects the trait that produces greater reproductive fitness, (RS4) seems strong. There is, however, as before, the problem of whether the best trait (in terms of reproductive fitness) is *available*. It may be that the cognitive mechanism that would lead to the greatest reproductive advantage compared to its competitors has yet to evolve and, as a result, is not available for natural selection to choose. As before with respect to (4), there are problems with the move to modify (RS4) as follows:

> (RS4') Biological evolution favors heuristics *from among those that are available* that lead to the greatest reproductive success.

Although this premise, by design, is immune to the availability objection, for the argument to go through with this modified premise, (RS5) needs to be modified as follows:

> (RS5') A rational heuristic is one that leads to the greatest reproductive success *among those heuristics that are available*.

The problem with this premise—as I argued in section VII above with respect to (5'), the claim that the most truth-tropic of the available heuristics is a rational one—is that rationality is not relative to biological factors such as what heuristics are available to a particular species.

Supposing, however, that the availability objection can be answered, perhaps because there are good reasons to believe that the best cognitive mechanisms in terms of reproductive success are in fact available. Even so, the claim that a rational heuristic is one that leads to the greatest reproductive success, (RS5) or (RS5'), is not as plausible as it first seemed. Recall the sense of rational that is of interest to friends and foes of the rationality thesis: humans are rational if they reason in accordance with the norms of reasoning. The sense of rational involved in (RS5) is not normative in the same way because we can imagine the heuristic that leads to the greatest reproductive success diverging from the norms of reasoning. It might be that making *plausibility* judgments in situations where *probability* judgments are called for leads to the greatest reproductive success (perhaps because plausibility judgments can be made more quickly), but, even so, it is not *rational* to make plausibility judgments when probability judgments are called for (see my discussion of plausibility versus probability in section III above). So, (RS5) and (RS5') are false—a heuristic that leads to the greatest reproductive success may well not be rational.[74] The attempt to save the evolutionary argument by linking evolution to rationality via reproductive success thus fails.

IX.

Suppose that contrary to the argument I sketched in section II above, human cognitive mechanisms are not primarily innate but are instead mostly learned. If this is right, the argument for the rationality thesis based on biological evolution is—even if all the objections I have considered are set aside—a non-starter. Still, if, as some have argued, the development of knowledge fits the general model of natural selection, an argument for the rationality thesis might be salvageable, such an argument, because it does not involve biological evolution, might be immune to the objections I have raised to the versions of the evolutionary argument considered so far. The idea of using natural selection as a model for the development of knowledge—sometimes called conceptual evolution—goes under the name of evolutionary epistemology. As part of articulating the conceptual evolution version of the evolutionary argument for the rationality thesis, I shall turn to a description of evolutionary epistemology.

[74] This argument against (RS5) or (RS5') is even stronger if reproductive success is understood in more precise terms. Strictly speaking, natural selection is driven by reproductive success from the point of view of genes. Organisms, to borrow a phrase from *The Selfish Gene*, are merely "survival machines" for genes. But if reproductive success is measured from the *genetic* point of view, it is even less likely that a heuristic that leads to the greatest reproductive success will be rational (in the normative sense) for humans.

Evolutionary epistemology is an approach to the theory of knowledge that sees a significant similarity between the growth of knowledge and biological evolution. An evolutionary epistemologist claims that the development of human knowledge proceeds through some natural selection process analogous to Darwin's theory of biological natural selection. The three major components of the model of natural selection are variation, selection, and retention. According to Darwin's theory of natural selection, variations are not predesigned to perform certain functions. Rather, those variations that perform useful functions are selected while those that do not are not selected; such selection is responsible for the appearance that variations intentionally occur. In the modern theory of evolution, genetic mutations provide the blind variations (blind in the sense that variations are not influenced by the effects they would have—the likelihood of a mutation is not correlated with the benefits or liabilities that mutation would confer on the organism), the environment provides the filter of selection, and reproduction provides the retention. Fit is achieved because those organisms with features that make them less fit for survival do not survive in competition with other organisms in the environment that have more fit features. Evolutionary epistemology applies this blind variation and selective retention model to the growth of scientific knowledge and to human thought processes in general. According to this view,[75] the development of human knowledge is governed by a process analogous to biological natural selection, rather than by an instance of the mechanism itself.[76] This

[75] This view is called the "evolution of theories program" (EET) by Bradie, "Assessing Evolutionary Epistemology," *Biology and Philosophy* 1 (1986), pp. 401-459.

[76] There is another view that gets called evolutionary epistemology; it is the *literal* version of evolutionary epistemology because it sees biological evolution as the primary cause of the growth of knowledge. This view dovetails nicely with the claim that our cognitive mechanisms are innate. On this view, called the "evolution of cognitive mechanisms program" (EEM) in "Assessing Evolutionary Epistemology," the growth of knowledge occurs through blind variation and selective retention because biological natural selection itself is the cause of conceptual variation and selection. A plausible version of the literal view need not hold that all human *beliefs* are innate but rather can hold that the mental mechanisms that guide the acquisition of non-innate beliefs are themselves innate and the result of biological natural selection. This version of evolutionary epistemology is not directly relevant to the argument for the rationality thesis currently under consideration. Ruse, *Taking Darwin Seriously*, chapter 5, defends a version of literal evolutionary epistemology which he links to sociobiology. See "Assessing" and Bradie, "Should Epistemologists Take Darwin Seriously?," in Nicholas Rescher, ed., *Evolution, Cognition and Realism* (Lanham, MD: University Press of America, 1990), pp. 33-38, for criticism of this view.

version of evolutionary epistemology, introduced and elaborated by Donald Campbell and Karl Popper,[77] sees the (partial) fit between theories and the world as explained by a mental process of trial and error. Roughly, evolutionary epistemology explains the development of knowledge through a process of the survival of the best (that is, most fit) beliefs.

With this picture of evolutionary epistemology in hand and recalling the original appeal of the evolutionary argument for rationality, it is easy to see the structure and strength of the evolutionary epistemological argument for the rationality thesis. The straightforward version of the evolutionary argument for rationality attempts to forge a link between biological evolution and rationality via truth but this failed because biological evolution cannot deliver truth; if evolutionary epistemology provides the right account of the origin of beliefs, that is, if human knowledge develops by selecting for the best beliefs, then there is a connection between evolution (in the general, *not* the biological, sense of the term) and truth, and thus hope for an evolutionary argument for rationality.

The argument can be laid out as follows:

(EE4) Conceptual evolution favors heuristics that are good at producing true beliefs.
(EE5) A rational heuristic is one that is good at producing true beliefs.
(EE6) Therefore, conceptual evolution favors rational heuristics.
(EE7) Humans follow heuristics which are the result of conceptual evolution.
(EE8) Therefore, there is a strong reason to believe that humans are rational—in particular, there is a strong reason not to interpret psychological experiments as demonstrating human irrationality.

This argument is not open to some of the same sorts of objections that counted against the original evolutionary argument, both because it is independent of the biological details of

[77] Donald Campbell, "Evolutionary Epistemology," *The Philosophy of Karl Popper*, book 1, P.A. Schilpp, ed. (LaSalle, IL: Open Court, 1974), pp. 413-463, reprinted in *Evolution, Theory of Rationality and the Sociology of Knowledge*, G. Radnitsky and W.W. Bartley III, eds. (LaSalle, IL: Open Court, 1987), pp.47-89; "Unjustified Variation and Selection in Scientific Discovery," *Studies in Philosophy of Biology*, F.J. Ayala and T. Dobzhansky, eds. (Berkeley: University of California Press, 1974), pp. 139-161; Karl Popper, *Conjectures and Refutations* (NY: Basic Books, 1962); *Objective Knowledge* (Oxford: Oxford University Press, 1972); and "Natural Selection and the Emergence of Mind," *Dialectica* 32 (1978), pp. 339-355, reprinted in *Evolution, Theory of Rationality and the Sociology of Knowledge*, pp. 139-156. See also W. Callebaut and R. Pinxter, eds., *Evolutionary Epistemology: A Multiparadigm Program* (Dordrecht, The Netherlands: D. Reidel Publishing Co., 1987), which includes an extensive bibliography.

our beliefs and because the availability objection seems to be less of a problem. These worries aside, is (EE4) plausible? First, (EE4)'s plausibility rests partly on the plausibility of evolutionary epistemology, a philosophical program that has come under much scrutiny.[78] Assessing the various criticism of it are, however, beyond the scope of this project. The second question, which *is* within the scope of this project, is whether conceptual evolution is, as some evolutionary epistemologists claim, a reliable producer of truth. I shall ask this question in terms of whether evolutionary epistemology is compatible, as it seems initially to be, with convergent realism, the view that as human beliefs change they make progress towards the truth. This will determine if (EE4) is true since conceptual evolution can only be truth-tropic if our beliefs get closer to the truth as they develop.[79]

I begin with a question about biology: does biological evolution progress towards a goal? Early theories of the origin of species included the belief in a chain of being, a hierarchy of living things ranging from the most primitive to the most advanced. In its early incarnation, the chain of being was believed to be static and to have been created by a god. In its later conceptions, the chain of being hierarchy was seen as pre-planned but gradually unfolding through time.[80] This picture of a temporalized chain of being sees biological progress as heading towards a final, particular and attainable goal. Darwin was responsible for the shift away from the belief in a divine and teleological view of the origin of species. Two of the most revolutionary insights of his theory were its move away from

[78] For critiques of evolutionary epistemology, see Peter Skagestad, "Taking Evolution Seriously: Critical Comments on D.T. Campbell's Evolutionary Epistemology," *Monist* 61 (October 1978), pp. 611-621; *Taking Darwin Seriously*, chapter two; Paul Thagard, "Against Evolutionary Epistemology," in P. D. Asquith and R. N. Giere, eds., *PSA 1980* (1980) pp. 187-196; and "Assessing Evolutionary Epistemology." Edward Stein and Peter Lipton, "Where Guesses Come From: Evolutionary Epistemology and the Anomaly of Guided Variation," *Biology and Philosophy* 4 (1990), pp. 33-56, attempt to answer some of these criticisms.

[79] The following discussion is adapted from my paper "Getting Closer to the Truth: Realism and the Metaphysical and Epistemological Ramifications of Evolutionary Epistemology," in *Evolution, Cognition and Realism*, pp. 119-129.

[80] The best known versions of this temporal chain of being view are those defended by Charles Bonnet and Jean-Baptiste Robinet in the middle of the eighteenth century.

teleology (the view that the evolution of species progresses towards an ultimate goal) and its elimination of a designer without the elimination of design. Darwin argued that speciation could occur by natural selection through time: it does not require a goal or the design of an all-powerful creator. Darwin saw evolution as a branching process: an original species gave rise to a variety of organisms within the species, which in turn gave rise, through speciation, to further distinct species. He did, however, want to preserve the notion that humans are more advanced since they appear later in the branching process. Today, we agree with Darwin in viewing evolution as non-teleological and independent of a designer. We also share with Darwin the intuitive sense that, in some way, we are more advanced than other species. But, in light of modern biology, this last belief has been cast in doubt—it is far from clear that in any qualitative, biological sense we are more advanced than other species.[81] What is clear is that, contrary to the chain of being view, there need be no ideal species towards which evolution progresses—in fact, progress towards an ideal species is inconsistent with contemporary evolutionary biology.

With this discussion as background, I will now turn to conceptual evolution. Whereas in the case of biological evolution we postulate no goal to progress towards, *prima facie* there does seem to be a goal for conceptual evolution, namely truth. It seems, for example, that scientific knowledge progresses towards a correct theory of the world. If so, this fact gives rise to a dramatic disanalogy between biological and conceptual evolution.[82]

I will now offer two arguments that evolutionary epistemology is not consistent with convergent realism. The first argument, which I call the "no goal-directed progress" argument, says that conceptual evolution, if it is to be analogous to biological evolution,

[81] See Francisco Ayala, "The Concept of Biological Progress," in *Studies in the Philosophy of Biology*, F.J. Ayala and T. Dobzhansky, eds. (Berkeley: University of California Press, 1974), for further discussion of this point.

[82] In spite of the threat of disanalogy, many evolutionary epistemologists hold the view that science does, in fact, progress towards the truth. For example, both Donald Campbell and Karl Popper, while they realize that there is great potential for conflict with their versions of evolutionary epistemology, seem to believe in convergent realism—even if science can never get to the truth, it can at least approach it. See "Evolutionary Epistemology," pp. 447-448, and *Objective Knowledge*, p. 58.

cannot get closer to the truth because there is no goal in biological evolution. The force of this argument can be avoided if, as seems reasonable, some disanalogy between biological and conceptual evolution is allowed. The second argument, which I call the "veil of selection" argument, says that conceptual evolution cannot progress towards the truth because the criterion of selection, since it is pragmatic, is too broad. This argument, counts against (EE4) since it shows that conceptual evolution cannot be relied upon to produce truth-tropic heuristics.

An essential feature of biological evolution is that it has no goal to approach, there is no ideal state towards which it progresses. An epistemology modeled on biological evolution cannot progress towards a goal if the analogy is to hold. If no goal is approached in epistemology, then both convergent realism and (EE4) are false since both require progress towards truth. Call this the "no goal-directed progress" argument.

A response to this argument would be that the mere absence of progress towards a goal does not mean there can be no progress. This much I agree with, but I think that there cannot be progress *towards truth* in an evolutionary epistemology, and that is what is at issue in (EE4). There are certain types of progress that can occur without a goal, while there are others that cannot—some processes require a goal in order to have a metric of progress, while others do not. The distinction I am drawing is between convergent progress and non-convergent progress. Convergent progress involves getting closer to a goal while non-convergent progress involves improvement but without a goal. The sort of progress that involves truth is convergent progress; the only possible metric for getting closer to the truth requires a complete true account of the world to get closer to.[83] The "no goal-directed progress" argument thus counts against (EE4).

[83] Note that giving up on progress towards truth does not entail giving up on progress completely. For example, Thomas Kuhn, *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1970)—see pp. 172-3, for his classic formulation—and Larry Laudan, *Progress and Its Problems* (Berkeley: University of California Press, 1977) reject convergent realism, but do not thereby reject every sense of progress. Kuhn removes the notion of a fixed truth from science, replacing it with an ***environmentally dependent*** notion of optimality, a notion that includes "maximum accuracy of predictions, degrees of specialization and number of concrete problem solutions," "Reflections on My

There is, however, a current trend towards a type of evolutionary epistemology called *natural selection* epistemology that might defuse this argument. Some evolutionary epistemologists, rather than arguing for a close analogy with biology, argue that change in both epistemology and biology are distinct instances of selection processes.[84] A natural selection epistemology thus need not match biological evolution in every feature so long as it exemplifies "general selection theory." A natural selection epistemologist could, thus, respond to the "no goal-directed progress" argument by saying that there are at least two sorts of selection processes, those that have a goal and those that do not—biological evolution does not have a goal but conceptual evolution does—and that while both sorts of evolutionary processes fit the model of natural selection, they differ in at least this one way. The move to a natural selection epistemology may save (EE4) and the evolutionary epistemologist's attempt to embrace convergent realism, but, even if it does, it does so only momentarily.

Critics," in *Criticism and the Growth of Knowledge*, Imre Lakatos and Alan Musgrave, eds. (London: Cambridge University Press, 1970), p. 264. See also Kuhn's "Logic of Discovery or Psychology of Research," in *The Essential Tension* (Chicago: University of Chicago Press, 1977), especially p. 289. While Kuhn and Laudan do not think that the development of scientific knowledge progresses *towards truth*, they do see it as progressing in some sense. I agree; progress is directional change towards a better state, (see "The Concept of Biological Progress" for a detailed discussion of this definition) but knowing the truth is not the only possible better state—the development of theories could progress by increasing the number of phenomena explained, the number of problems solved, the number of bridges built, etc. Although work needs to be done to develop a precise notion of scientific progress to take the place of getting closer to the truth, I think that strides have been made towards a (non-truth-tropic) notion of progress suited to evolutionary epistemology. I am not, however, as confident as Kuhn, (for example, see "Reflections on My Critics," p. 264), is that it would be easy to develop such a notion. Consider, for example, how difficult it is to talk about progress in biological evolution, where the mechanisms of variation and selection are far better characterized and understood than in conceptual evolution. Thus while the "no goal-directed progress" argument allows for progress in conceptual evolution, it does not allow for progress towards the truth.

[84] See, for example, Donald Campbell, "Selection Theory and the Sociology of Scientific Validity," in *Evolutionary Epistemology: A Multiparadigm Program*; Donald Campbell and Bonnie Paller, "Extending Evolutionary Epistemology to 'Justifying' Scientific Beliefs," in *Issues in Evolutionary Epistemology*, K. Hahlweg and C.A. Hooker, eds. (Albany: State University of New York Press, 1989); and David Hull, "A Mechanism and Its Metaphysics," *Biology and Philosophy* 3 (April 1988), pp. 123-155. I have my doubts about the move to a natural selection epistemology. If the notion of selection theory is made too much broader it will include almost every developmental process. Broadening the notion of selection theory is supposed to save evolutionary epistemology from the force of various disanalogies that some have argued hold between conceptual and biological evolution. It seems to me that these disanalogies are too quickly admitted by many evolutionary epistemologists; rather than embracing these disanalogies, many of them should be attacked head on. See "Where Guesses Come From," *Biology and Philosophy* 4 (1989), pp. 33-56, for an attempt to address some of the supposed disanalogies between biological and conceptual evolution.

I will now proceed to my second argument that evolutionary epistemology is incompatible with convergent realism and hence that (EE4) is false. Evolutionary epistemology attempts (as does natural selection epistemology) to apply the natural selection model of blind variation and selective retention to conceptual evolution. Put simply, in biology (focusing on just the selection part of the process), the organism that fits better in the environment than its competitors is the organism that is selected. This does not mean the organism is *ideally* suited to that particular environment—only that it is well enough suited to survive and better suited than its competitors. The same sort of account is given to explain why a belief or scientific theory is selected—the belief selected is better fit than its competitors. Here "better fit" amounts to better behavioral consequences or closer correlation with the data. As a result according to evolutionary epistemology, selection cannot distinguish between knowledge and mere useful opinion, that is, it cannot distinguish between (locally) optimal behavior based on truths and equally optimal behavior based on falsehoods. If we read "better fit" to mean better fit with the data, the same point can be made by appealing to underdetermination of theory by data: the data are consistent with many more theories (beliefs) than just the true one. In other words, there is a "veil of selection" in evolutionary epistemology—the selection criterion of optimality fails to guarantee that true theories will be selected. If we think of selection as a filter, in the case of evolutionary epistemology the filter is too porous to precipitate only true theories—other theories will pass through the filter as well. Since it holds that belief selection is neutral with respect to truth (that is, selection is for good results not for truth), evolutionary epistemology is not consistent with guaranteed progress towards truth. Selection is based on optimality but, since there are many false beliefs that will work just as well as any true one, optimality does not amount to truth. The conclusion of the veil of selection argument is that since selection in evolutionary epistemology is based on pragmatic criteria, it cannot guarantee progress towards the truth.

The "no goal" and the "veil of selection" arguments count against (EE4), the evolutionary epistemological claim that our beliefs will get closer to the truth as they develop (in other words, as they "evolve"). This undermines the evolutionary epistemology argument for the rationality thesis by blocking the link between evolution and truth that seemed a promising route to connect evolution to rationality. The evolutionary epistemology argument for the rationality thesis thus fails.

X.

In the first several sections of this chapter, I argued that biological evolution cannot be relied upon to produce heuristics that are truth-tropic. In the most recent section, I argued that the same is true for conceptual evolution. That conceptual evolution fails to produce true beliefs may seem unconnected to why biological evolution fails to do so, since biological evolution fails to do so for reasons that seemed connected with the details of biological evolution. Conceptual evolution, however, fails to produce true beliefs because the filter of natural selection is not fine enough to select for truth (or rationality). These reasons are similar to why biological evolution fails to produce true beliefs. This similarity indicates that there is a deeper reason—one having to do with the abstract model of natural selection—why the evolutionary argument for rationality, in both its biological and evolutionary epistemological forms, will not work. This feature is, as emphasized above, that the filter of selection (whether natural or otherwise) is designed to precipitate pragmatic results, and not truth. This shows that evolutionary arguments, whatever form they take, will not suffice to defend the rationality thesis.

I began this chapter by comparing the mechanisms of cognition to the mechanisms of vision. I suggested that just as natural selection has done an excellent job selecting a visual system that helps us generate true beliefs about the world, sc natural selection should be expected to have selected a cognitive system that generates true beliefs about the world. Much of the burden of this chapter has been to show that natural selection cannot, in fact, be relied upon to do so with respect to cognition. This might indicate that something is wrong with my comparison between cognition and vision. Actually, I think the comparison was apt but the characterization of both cognition and vision and their relation to natural selection was mistaken. I originally characterized the visual system as doing an excellent job of generating true beliefs about the world. Although this may well be true, it is a mistake to confuse this with the claim that the visual system was selected *for* generating true beliefs. In fact, the visual system, like our cognitive mechanisms, was selected for because of its pragmatic effects. If the visual system is good at producing true beliefs, this is a (happy) side effect of its contribution to the selective advantageousness of genes for eyes, the visual cortex, etc., not because natural selection has selected *for* truth-tropic visual mechanisms. Thus, the ramifications of this chapter go beyond the evolutionary argument for the rationality thesis; they show that the attempt to connect natural selection to truth or rationality in any realm—not just cognition—will fail.

Chapter Six: Conclusion

I.

I began Chapter One by noting that it has become a commonplace observation that humans are irrational. Cognitive scientists, armed with the results of the irrationality experiments, have precisely formulated this commonplace observation and provided it with what seems to be strong empirical support. Once the irrationality thesis has been precisely formulated, it becomes vulnerable to some creative and seemingly plausible arguments to the effect that cognitive science cannot possibly prove that humans are irrational. In the preceding chapters, I have considered such arguments for the rationality thesis. None of these arguments dispute that cognitive science is the science of the mind and its operation. Many of them (namely, all of the arguments I have considered with the exception of the argument based on the misinterpretation strategy), however, imply that cognitive science cannot assert the thesis that humans are irrational because, for reasons outside of cognitive science, humans must be rational. Having assessed these several arguments for the rationality thesis, I return to the question of the limits of cognitive science.

It may seem strange at first that arguments for the rationality thesis imply that it is not the business of cognitive science to make certain claims about human rationality, but this initial counter-intuitive aura disappears in the light of comparison with other cases where there are theoretical limits on science. Recall my example in Chapter One involving the claim that it is not the business of biology to say what is moral. In particular, consider the claim that some biological facts determine that incest is morally wrong.[1] Even if a biologist discovered that humans have an innate belief that incest is wrong and/or that any offspring resulting from incestuous copulation is highly likely to have serious genetic defects, the

[1] I pick incest because it is a favorite example of sociobiologists. See, for example, E. O. Wilson, *Sociobiology: The New Synthesis* (Cambridge: Harvard University Press, 1975), pp. 78-79; *On Human Nature* (Cambridge: Harvard University Press, 1978), pp. 37-40; Charles Lumsden and E. O. Wilson, *Genes, Minds, and Culture* (Cambridge: Harvard University Press, 1981), pp. 85-86, 147-158; *Promethean Fire* (Cambridge: Harvard University Press, 1983), pp. 125-138, 175-180; and Michael Ruse, *Taking Darwin Seriously* (Oxford: Basil Blackwell, 1986), pp. 145-147.

moral wrongness of incest would not thereby be established. Biological facts do not determine moral facts. Biology is thus limited in the claims it can make about human morality. In light of the fact that another scientific discipline is limited in the normative claims it can make, it should be less surprising if cognitive science is limited in the claims it can make about rationality.

Of the arguments considered in the previous chapters, not one was successful in establishing the rationality thesis and thus implying that cognitive science is limited with respect to what it can say about human rationality. This does not necessarily mean that cognitive science is unlimited in what it can say about rationality any more than the failure of (philosophical or biological) arguments to establish that humans are moral means that biology is not limited in what it can say about morality. Cognitive science may still have limits even if the rationality thesis is false. Although I cannot fully support this contention based on the arguments I have made in the preceding chapters, I do want to suggest a way in which cognitive science is limited.

The question "Are humans rational?" involves two quite separate questions: "What are the norms of reasoning?" and "Do humans match these norms?" Cognitive science is seems relevant to the latter question—if the normative standards of reasoning have been determined, then it may be the business of cognitive science to determine whether humans reason in accordance with these standards. But it is a separate question whether cognitive science gets to play a role (and, if so, what kind of role) in determining what these normative standards are.

In Chapter Three, I discussed the wide reflective equilibrium account of how the norms of reasoning are justified. Roughly, the wide reflective equilibrium process counts as justified the principles that result from the balancing of (a) first-order intuitions about what particular inferences are acceptable, (b) second-order intuitions about what, in general, counts as good reasoning, and (c) general theoretical and philosophical considerations. I dwelled on one serious objection to this account, namely, that wide reflective equilibrium counts as justified principles that might, from an objective point of view, be irrational. I then sketched two responses to this objection. The first response was that wide reflective equilibrium produces the maximally coherent set of principles of reasoning, and all it is for a principle to be rational is for it to be in the maximally coherent set of principles. The second response was vaguer, but, I think more promising. This response involved arguing that wide reflective equilibrium is perhaps the best that can be done—one may be asking too much of a theory of justification if one requires that it deems justified all and only those principles that are objectively rational. Now it seems to me that this second response is on the right track. This is not, however, the place to mount a complete defense of the wide reflective equilibrium account of justifying principles of reasoning. I do, however, want to sketch a weaker claim about the process of justifying such principles.

Recall from Chapter Two my claim that the norms of reasoning are not indexed to human cognitive competence (in contrast to the way that the norms of linguistics are indexed to linguistic competence). This claim suggests that whatever the process of justifying the norms of reasoning is, the process will involve considerations that extend beyond the details of the human mind (brain), in particular, beyond human cognitive competence. But, to the extent that determining the normative principles of reasoning requires considerations that extend beyond empirical facts about the human mind, determining the principles of reasoning is outside the realm of cognitive science. As such, the relationship between cognitive science and rationality is like the relationship between biology and morality and *unlike* the relationship between linguistics and grammaticality.

The question "Do humans speak grammatically?" can be separated into two questions: "What are the norms of grammaticality?" and "Do humans match these norms?" As I argued in Chapter Two, the norms of grammaticality are indexed to actual linguistic competence; as such, the empirical inquiry of linguistics is relevant to determining the norms. Further, setting aside performance errors, whether humans match these norms of grammaticality is not an empirical question once the norms have been determined; since the norms are indexed to actual linguistic competence, they necessarily match; in linguistics, the normative principles just are the principles that characterize linguistic competence. The question "Are humans moral?" is, in contrast to the linguistics case, like the question "Are humans rational?" The question about morality involves the questions "What are the norms of morality?" and "Do humans follow these norms?" The second question involves empirical issues in biology (and/or psychology), while the former question does not. Note that this parallels what I had to say about wide reflective equilibrium at the end of Chapter Three; to the extent that they fit the reflective equilibrium model morality and rationality involve *wide* reflective equilibrium, while grammaticality involves *narrow* reflective equilibrium.

This analysis of the question "Are humans rational?" does entail that there are some limits on cognitive science, namely that cognitive science does not determine what the norms of reasoning are. While this is a limitation on cognitive science, it is not the sort of limit that is implied by the arguments for the rationality thesis considered in the previous four chapters. These arguments for the rationality thesis attempted to show that either for conceptual reasons (as in the reflective equilibrium argument or the principle of charity argument) or empirical reasons (as in the evolutionary argument or the argument based on the misinterpretation strategy), the answer to the question "Do humans match the norms of reasoning?" has to be "yes." The conceptual arguments for the rationality thesis have the further implication that the details of cognitive science are completely irrelevant to determining whether or not humans are rational. These arguments failed to demonstrate that humans must be rational and thereby they failed to demonstrate that cognitive science is limited in what it can show about human reasoning.

II.

Given that cognitive science *can* possibly show that humans are irrational, do the irrationality experiments that I have been considering in fact show that humans are irrational? At least on the surface, these experiments seem like the right sort to show that human reasoning diverges from the normative standards of reasoning. For example, the results of the conjunction experiment seem to show that humans systematically ignore the conjunction rule in probability theory; they are thereby violating a norm of reasoning, and hence seem to be irrational. Matters are not, however, so simple for two related reasons. First, it is possible that some feature of the experimental design is flawed in such a way that the errors that appear do not represent actual cases where human cognitive competence diverges from the norms. This is the claim Gigerenzer makes with respect to the conjunction experiment, and it is the general sort of claim championed by friends of the misinterpretation strategy argument for the irrationality thesis. In Chapter Four, I reviewed the argument that the results of the conjunction experiment are due to the way the task is phrased; if the conjunction experiment is phrased in terms of frequency rather than probability (that is, how many people out of one hundred who fit the Linda description will be bank tellers rather than what is the probability that Linda is a bank tellers), subjects respond in accordance with the norms.[2] Also in Chapter Four, I noted the incompleteness of this argument as a defense of the rationality thesis: showing that subjects do not make systematic errors related to frequencies does not, without further explanation, explain away the errors subjects make on the probability task. Such a further explanation would need to demonstrate that the errors subjects make on the probability version of the conjunction task are not errors caused by having a cognitive competence that diverges from the norms. I have argued that none of the suggested conceptual arguments for the rationality thesis will suffice to show that human cognitive competence must not diverge from the norms. However, although the misinterpretation strategy does not work as a general argument for the rationality thesis, this strategy might be well-founded in particular cases.

[2] Gerd Gigerenzer, "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases,'" *European Review of Social Psychology* 2 (1991), pp. 83-115.

The second point why the irrationality experiments do not simply prove humans are irrational is because human reasoning might diverge from the norms of reasoning due to performance errors. Although no successful conceptual argument has been given that demonstrates that all divergences from the norms of reasoning are due to performance errors, no argument has been given that some, most or even all of such divergences are *not* due to performance errors. The irrationality thesis will not be established without showing that *some* of the divergences from the norms are in fact due to the structure of cognitive competence and not to performance errors. It is no simple task to show that an instance of divergence from the norms is really due to cognitive competence and not a performance error. This is true particularly if a strong version of the weak principle of charity (for example, one that says humans should be interpreted as rational unless there is strong empirical evidence to the contrary) is true.

The upshot of these points is that cognitive competence is epistemologically opaque. Human cognitive competence cannot simply be read off of behavior (because, for example, that behavior may be due to performance errors). Further, cognitive competence cannot simply be read off of the neurological structure of the brain (for the reasons discussed in Chapter Four, section II). The situation is that we know what would settle the issue between the rationality and the irrationality thesis—it is the set of empirical facts about human cognitive competence and whether this competence matches certain principles (the norms of reasoning)—but these facts are difficult to discover.

Chomsky has something to say about this problem as it relates to discovering and understanding linguistic competence. He says that among the main tasks of the linguist are providing answers to the questions "What is a system of knowledge?," "What is in the mind/brain of the speaker of English or Spanish or Japanese?," "How does this system of knowledge arise in the mind/brain?," and "How is this knowledge put to use in speech?"[3] As the linguist begins to have answers to these questions,

> the brain scientist can begin to explore the physical mechanisms that exhibit the properties revealed in the linguist's abstract theory. In the absence of answers to these questions, brain scientists do not know what they are searching for; their inquiry is in this respect blind.
>
> This is a familiar story in the physical sciences. Thus nineteenth-century chemistry was concerned with the properties of chemical elements and provided models of compounds (for example, the benzene ring). It developed such notions as valence, molecule, and the periodic table of the elements. All of this proceeded at a level that was highly abstract. How all of this might relate to more fundamental physical mechanisms was unknown, and there was in fact much debate over whether these notions had any "physical reality" or were just convenient myths devised to help organize experience. This abstract inquiry set problems for the physicist: to discover physical mechanisms that exhibit these properties. The remarkable successes of twentieth-century physics have provided increasingly more sophisticated and compelling solutions for these problems in a quest that some feel may be approaching a kind of "ultimate and complete answer."
>
> The study of the mind/brain today can be usefully conceived in much the same terms. When we speak of the mind, we are speaking at some level of abstraction of yet-unknown physical mechanisms of the brain, much as those who spoke of the valence of oxygen or the benzene ring were speaking at some level of abstraction about physical mechanisms, then unknown. Just as the discoveries of the chemist set the stage for further inquiry into underlying mechanisms, so today the discoveries of the linguist-psychologist set the stage for further inquiry into brain mechanisms, inquiry that must proceed blindly, without knowing what it is looking for, in the absence of such understanding, expressed at an abstract level.[4]

The same sort of comments seem applicable to attempts of the cognitive scientist and the brain scientist to discover, describe and understand human cognitive competence. The cognitive scientist must try to understand mechanisms of reasoning at an abstract level, (that is, at the mental level) without well-developed knowledge of the neurological structures (whether primarily innate or primarily acquired through interaction with the environment) that embody them. Only with guidance from cognitive science—guidance based on these sort of abstract explanations—can the brain scientist undertake an informed inquiry into the neurological features of the brain responsible for cognitive competence.

[3] Noam Chomsky, *Language and the Problems of Knowledge: The Managua Lectures* (Cambridge: MIT Press, 1988), p. 4

[4] *Ibid.*, pp. 6-7.

This inquiry, in turn, can influence the work of the cognitive scientist; by better understanding the underlying neurological str ıctures, the cognitive scientist can better characterize the mental mechanisms involved in human cognitive competence.

Friends of sociobiological explanations might object at this point that the account of the task of the cognitive scientist that I give here is actually more appropriately applied to biologists interested in explaining human behavior. To develop an accurate characterization of human cognitive competence, one must understand the ecological circumstances in which such a competence evolved, was selected, etc., (if human cognitive competence is primarily innate) or in which the capacity to develop such a competence (through learning) evolved, was selected, etc. (if human cognitive competence is *not* primarily innate). It is, such an objection might continue, the sociobiologist, not the cognitive scientist, who is ideally positioned to undertake such an investigation.

There is something right about this point, though there is something wrong about it as well. Understanding the various selective pressures that have been involved in the development of human mental mechanisms including cognitive competence could well be useful to the project of the cognitive scientist as I describe it above. With only the knowledge of human behavior, cognitive scientists need whatever guidance they can get as to what cognitive competence is. Evolutionary considerations seem an appropriate place to look for such guidance. A paradigmatic example of drawing on evolutionary considerations is Cosmides's work on the selection task as discussed in Chapter One.[5] As a clue to discovering the nature of the heuristics people are following in the selection task, Cosmides turned to evolutionary theory for guidance about the sort of innate heuristics (what she calls "Darwinian algorithms") that would be adaptive given the environment and the information processing problems humans faced. According to Cosmides, evolutionary considerations suggest in particular that tasks having to do with social exchange would be

[5] Leda Cosmides, "The Logic of Selection: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task," *Cognition* 31 (1989), pp. 187-276.

particularly important problems for innate cognitive heuristics to address; heuristics that are specificaliy designed to solve social exchange tasks are thus likely to be operative in the selection task.

Just because evolutionary theory is occasionally drawn from not does not, however, make the inquiry into the nature of human cognitive competence a task for the sociobiologist. Sociobiologists typically make a direct leap from genes to behavior. So, for example, they might talk about natural selection selecting for certain reasoning behaviors. But this sort of talk clearly involves dramatic oversimplification. Natural selection involves selection for genes, but the connection between genes and behavior is quite complex; to say the least, the connection is mediated by the process of development (that is, the process through which genes, in the presence of certain environmental inputs, will produce embodied organisms) and by the mind (that is, genes cause certain mental mechanisms to develop which in turn engender certain sorts of behavior). Focusing on the role of the mind, the point is that what gets selected by natural selection are not behaviors, but rather a.e genes that lead to the development of the brain mechanisms that tend to cause behaviors. As Cosmides and Tooby put it, "Behavior cannot occur *sui generis*; behavior is an effect produced by . . . psychological mechanisms."[6] They argue, correctly, I think, that sociobiology is not the right approach for developing and understanding of mental mechanisms, but rather cognitive science informed by evolutionary considerations, what they call "evolutionary psychology,"[7] is.[8]

[6] Leda Cosmides and John Tooby, "From Evolution to Behavior: Evolutionary Psychology as the Missing Link," in *The Latest on the Best*, John Dupré, ed. (Cambridge: MIT Press, 1987), p. 282.

[7] See also, John Tooby, "The Emergence of Evolutionary Psychology," in *Emerging Syntheses in Science: Proceedings of the Founding Workshops of the Santa Fe Institute*," D. Pines, ed. (Redwood City, CA: Addison-Wesley, 1988), pp. 67-75; and John Tooby and Leda Cosmides, "Evolutionary Psychology and the Generation of Culture: Part I: Theoretical Considerations," *Ethology and Sociobiology* 10 (1987), pp. 29-49.

[8] Others who have endorsed evolutionary considerations as helpful in understanding mental mechanisms are David Marr, *Vision: A Computational Investigation into Human Representation and Processing of Visual Information* (San Francisco: Freeman, 1982), who argues that evolutionary considerations are relevant to developing an account of the visual mechanisms in the brain, and Steve Pinker and Paul Bloom, "Natural Language and Natural Selection," *Behavioral and Brain Science* 13 (1990), pp. 707-784, who cautiously suggest (for example, p. 726) that evolutionary considerations are relevant to our

Returning more specifically to the question of whether the irrationality experiments are relevant to the irrationality thesis, the question has now become whether such experiments can play a role in the abstract research program in which the cognitive scientist is engaged for the purpose of developing an account of human cognitive competence. The answer to this question seems obviously yes—human reasoning behavior is of course relevant to developing a theory of cognitive competence—but the problems of epistemological access to this competence discussed above show that these experiments need to be taken with a grain of salt, that is, they need to be taken in the context of (a) the possibility of performance errors, (b) evolutionary considerations, and (c) neurological and computational constraints (for example, constraints resulting from humans having brains of a particular size and humans being able to perform computations in a certain amount of time[9]). Given these various factors that ought to figure in the development of a theory of human cognitive competence, friends of the irrationality thesis may have spoken too soon when they took the results of the irrationality experiments to directly establish the irrationality thesis.

My conclusion thus does not really answer the question "Are humans rational?" Rather, my primary conclusion offers a clarification of what this question asks and an account of what sort of evidence would count towards answering this clarified question. The clarified question is whether humans have a cognitive competence that matches the (theoretically established) normative principles of reasoning. The sort of evidence required to answer this question varies; it includes neurological and computational constraints, evolutionary considerations, and behavioral evidence, in particular, evidence about the sort of behaviors involved in the irrationality experiments. The various arguments considered

understanding of human language. For more specific suggestions about evolutionary considerations that might be relevant to language, see some of the commentaries on Pinker and Bloom's article.

[9] For discussion of these sorts of constraints, see Christopher Cherniak, *Minimal Rationality* (Cambridge: MIT Press, 1986). Note that by agreeing with Cherniak that neurological and computational constraints are relevant to developing a theory of cognitive competence does not entail agreeing with him that such constraints are relevant to developing a normative account of rationality (minimal or otherwise).

in this essay that were supposed to answer this clarified question in the affirmative (that is, answer it by saying that humans *are* rational, that humans have a cognitive competence that matches the normative principles of reasoning) failed, but not all of them are necessarily irrelevant to the clarified question.

In Oscar Wilde's novel *The Picture of Dorian Gray*, the character of Lord Henry says, "I wonder who it was defined man as a rational animal. It was the most premature definition ever given."[10] My conclusion is somewhat similar to this. Although I have not concerned myself with those attempts to *define* humans as rational,[11] I have been concerned with those who attempt to *prove* that humans are rational. Their arguments are not premature; they are, I have argued, unsound. It is, however, their rivals, those who have tried to argue from the basis of the irrationality experiments to the conclusion that humans are irrational, who make arguments that may be premature; only great advances in cognitive science, advances that may be beyond the epistemological limits of humans, will settle the (clarified) question of whether humans are rational.

[10] Oscar Wilde, *The Picture of Dorian Gray* (London: Messrs Ward, Lock, and Company, 1891); reprinted (London: Penguin Books, 1985), p. 52 (reprinted edition).

[11] For example, Jonathan Bennett, *Rationality* (London: Routledge and Kegan Paul, 1964). See my discussion in Chapter One, section III.

BIBLIOGRAPHY

Atran, Scott, *Cognitive Foundations of Natural History* (Cambridge: Cambridge University Press, 1990).

Ayala, Francisco, "The Concept of Biological Progress," in *Studies in the Philosophy of Biology*, F.J. Ayala and T. Dobzhansky, eds. (Berkeley: University of California Press, 1974).

Berwick, Robert and Amy Weinberg, *The Grammatical Basis of Linguistic Performance* (Cambridge: MIT Press, 1980).

Block, Ned, "Advertisement for a Semantics for Psychology," in *Midwest Studies in Philosophy*, volume 10, P. A. French, et al., eds. (Minneapolis: University of Minnesota Press, 1986), pp. 615-677.

Braine, M. D. S., B. J. Resier, and B. Rumain, "Some Empirical Justification for a Theory of Natural Propositional Logic," in *The Psychology of Learning and Motivation*, volume 18, Gordon H. Bower, ed. (Orlando, FL: Academic Press, 1984), pp. 313-371.

Bennett, Jonathan, *Rationality* (London: Routledge and Kegan Paul, 1964).

Bradie, Michael, "Assessing Evolutionary Epistemology," *Biology and Philosophy* 1 (1986), pp. 401-459.

_____, "Should Epistemologists Take Darwin Seriously?," in *Evolution, Cognition and Realism.*, Nicholas Rescher, ed. (Lanham, MD: University Press of America, 1990), pp. 33-38.

Callebaut, W., and R. Pinxter, eds., *Evolutionary Epistemology: A Multiparadigm Program* (Dordrecht, The Netherlands: D. Reidel Publishing Co., 1987).

Campbell, Donald,"Evolutionary Epistemology," in *The Philosophy of Karl Popper*, volume 1, Paul Arthur Schilpp, ed. (LaSalle, IL: Open Court, 1974), pp. 413-463. Reprinted in *Evolution, Theory of Rationality and the Sociology of Knowledge*, G. Radnitsky and W.W. Bartley III, eds. (LaSalle, IL: Open Court, 1987), pp.47-89.

_____, "Selection Theory and the Sociology of Scientific Validity," in *Evolutionary Epistemology: A Multiparadigm Program*, W. Callebaut and R. Pinxter, eds., pp. 139-158.

_____, "Unjustified Variation and Selection in Scientific Discovery," *Studies in Philosophy of Biology*, F. J. Ayala and T. Dobzhansky, eds. (Berkeley: University of California Press, 1974), pp. 139-161.

Campbell, Donald and Bonnie Paller, "Extending Evolutionary Epistemology to 'Justifying' Scientific Beliefs," in *Issues in Evolutionary Epistemology*, K. Hahlweg and C. A. Hooker, eds. (Albany: State University of New York Press, 1989), pp. 231-257.

Caplan, David and Nancy Hildebrandt, *Disorders of Syntactic Comprehension* (Cambridge: MIT Press, 1988).

Caramazza, Alfonso, John Hart, and Rita Berndt, "Category-Specific Naming Deficit Following Cerebral Infraction," *Nature* 316 (August 1985), pp. 439-440.

Cartwright, Nancy, *How the Laws of Physics Lie* (Oxford: Oxford University Press, 1983).

Cheng, P. W., and K. J. Holyoak, "On the Natural Selection of Reasoning Theories," *Cognition* 33 (1989), pp. 285-333.

_____, "Pragmatic Reasoning Schemas," *Cognitive Psychology* 17 (1985), pp. 391-416.

Cherniak, Christopher, "The Bounded Brain: Toward Quantitative Neuroanatomy," *Journal of Cognitive Neuroscience* 2 (1990), pp. 58-68.

_____, *Minimal Rationality* (Cambridge: MIT Press, 1986).

_____, "Undebuggability and Cognitive Science," *Communications of the Association for Computing Machinery* 31 (1988), pp. 402-412.

Chomsky, Noam, *Aspects of the Theory of Syntax* (Cambridge: MIT Press, 1965).

_____, *Knowledge of Language* (New York: Praeger, 1986).

_____, *Language and Problems of Knowledge* (Cambridge: MIT Press, 1980).

_____, *Reflections on Language* (New York: Random House, 1975).

_____, *Rules and Representations* (New York: Columbia University Press, 1980).

_____, "Rules and Representations," *Behavioral and Brain Sciences* 3 (1980), pp 1-61.

_____, *Syntactic Structures* (The Hague, The Netherlands: Mouton, 1957).

Cosmides, Leda, "The Logic of Selection: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task," *Cognition* 31 (1989), pp. 187-276.

Cosmides, Leda, and John Tooby, "From Evolution to Behavior: Evolutionary Psychology as the Missing Link," in *The Latest on the Best: Essays on Evolution and Optimality*, John Dupré, ed. (Cambridge: MIT Press, 1987), pp. 277-306.

Cohen, L. J., "Can Human Irrationality Be Experimentally Demonstrated?," *Behavioral and Brain Sciences* 4 (1981), pp. 317-331 and "Continuing Commentary" in volume 6 (1983), pp 487-533.

_____, *The Dialogue of Reason* (Oxford: Oxford University Press, 1986).

_____, "On the Psychology of Prediction: Whose is the Fallacy?," *Cognition* 7 (1979), pp. 385-407.

_____, "Whose is the Fallacy?: A Rejoinder to Daniel Kahneman and Amos Tversky," *Cognition* 8 (1980), pp. 89-92.

Conee, Earl, and Richard Feldman, "Stich and Nisbett on Justifying Inference Rules," *Philosophy of Science* 50 (1983), pp. 326-331.

Crow, J., "Genes That Violate Mendel's Rules," *Scientific American* 240:2 (1979), pp. 134-146.

Daniels, Norman, "On Some Methods of Ethics and Linguistics," *Philosophical Studies* 37 (1980), pp. 21-36.

_____,"Wide Reflective Equilibrium and Archimedean Points," *Canadian Journal of Philosophy* 10 (1980), pp. 83-103.

_____, "Wide Reflective Equilibrium and Theory Acceptance in Ethics," *Journal of Philosophy* 76 (1979), pp. 256-282.

Davidson, Donald, "Incoherence and Irrationality," *Dialectica* 39 (1985), pp. 345-354.

_____, *Inquiries into Truth and Interpretation* (Oxford: Oxford University Press, 1984).

Dawkins, Richard, *The Blind Watchmaker* (New York: W.W. Norton, 1986).

_____, *The Selfish Gene* (Oxford: Oxford University Press, 1976).

Dennett, Daniel, *Brainstorms* (Cambridge: MIT Press, 1978).

_____, "Cognitive Wheels: The Frame Problem in AI," in *Minds, Machines and Evolution*, Christopher Hookaway, ed. (Cambridge: Cambridge University Press, 1984), pp. 17-42.

_____, *The Intentional Stance* (Cambridge: MIT Press, 1987).

Feldman, Richard, "Rationality, Reliability and Natural Selection," *Philosophy of Science* 55 (1988), pp. 218-227.

Fiedler, Klaus, "The Dependence of the Conjunction Fallacy on Subtle Linguistic Factors," *Psychological Research* 50 (1988), pp. 123-129.

Flanagan, Owen, *The Science of the Mind* (Cambridge: MIT Press, 1984).

Fodor, Jerry, *The Modularity of Mind* (Cambridge: MIT Press, 1983).

_____, "Psychosemantics," in *Mind and Cognition: A Reader*, William Lycan, ed. (Oxford: Basil Blackwell, 1990), pp. 312-337.

_____, *Psychosemantics* (Cambridge: MIT Press, 1987).

_____, *A Theory of Content and Other Essays* (Cambridge: MIT Press, 1990).

_____, "Three Cheers for Propositional Attitudes," in *Representations* (Cambridge: MIT Press, 1981), pp. 100-123.

Fodor, Jerry, and Merrill Garrett, "Some Reflection on Competence and Performance," in *Psycholinguistic Papers*, J. Lyons and R. J. Wales, eds. (Edinburgh: Edinburgh University Press, 1966), pp. 135-154.

Gallistel, C. R., *The Organization of Learning* (Cambridge: MIT Press, 1990).

Garcia, J., McGowan, B. K., and Green, K. F., "Biological Constraints on Conditioning," in *Classical Conditioning: Current Research and Theory*, volume 2, A. H. Black and W.F. Prokasy, eds. (NY: Appleton-Century Crofts, 1972), pp. 3-27.

Garfield, Jay, "Review of *A Border Dispute*," *Journal of Symbolic Logic* 53 (1988), pp. 314-317.

Garfield, Jay, ed., *Modularity in Knowledge Representation and Natural-Language Understanding* (Cambridge: MIT Press, 1987)

Gigerenzer, Gerd, "How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases,'" *European Review of Social Psychology* 2 (1991), pp. 83-115.

_____, "On Cognitive Illusions and Rationality," in *Reasoning and Rationality: Essays in Honor of L. J. Cohen*, E. Eells and T. Maruszewski, eds. (Amsterdam: Rodopi, forthcoming).

Goldman, Alvin, *Epistemology and Cognition* (Cambridge: MIT Press, 1986).

Goldschmidt, Richard, *The Material Basis of Evolution* (New Haven: Yale University Press, 1940).

Goodman, Nelson, *Fact, Fiction and Forecast*, fourth edition (Cambridge: Harvard University Press, 1983).

Gould, Steven Jay, "The Limits of Adaptation: Is Language a Spandrel of the Human Brain?," talk presented to the Cognitive Science Seminar, Center for Cognitive Science, MIT (October 1987).

_____, "Panselectionist Pitfalls in Parker and Gibson's Model of the Evolution of Intelligence," *Behavioral and Brain Sciences* 2 (1979), pp. 385-386.

Gould, Steven Jay, and Richard Lewontin, "The Spandrels of San Marcos and the Panglossian Program: A Critique of the Adaptationist Programme," *Proceedings of the Royal Society of London* 205 (1978), pp. 281-288. Reprinted in *Conceptual Issues in Evolutionary Biology*, Elliott Sober, ed. (Cambridge: MIT Press, 1984), pp. 252-270.

Gould, Steven Jay, and E. S. Vrba, "Exaptation—a Missing Term in the Science of Form," *Paleobiology* 8 (1982), pp. 4-15.

Grandy, Richard, "Reference, Meaning and Belief," *Journal of Philosophy* 70 (1973), pp. 439-452.

Griffin, Donald, *Animal Thinking* (Cambridge: Harvard University Press, 1984).

Griggs, R. A., "The Role of Problem Content in the Wason Selection Task and THOG Problem," in *Thinking and Reasoning: Psychological Approaches*, Jonathan Evans, ed. (London: Routledge and Kegan Paul, 1983), pp. 16-43.

Heil, John, "Does Psychology Presuppose Rationality?," *Journal for the Theory of Social Behavior* 16 (1986), pp. 77-87.

Hernnstein, Richard, "Level of Stimulus Control: A Functional Approach," *Cognition* 37 (1990), pp. 133-166.

Henle, Mary, "On the Relation Between Logic and Thinking," *Psychological Review* 69 (1962), pp. 376-382.

_____, "Foreword," in *Human Reasoning*, R. Revlin and R. E. Mayer, ed. (Washington D.C.: Winston, 1978), pp. xiii-xviii.

Hull, David, "A Mechanism and Its Metaphysics," *Biology and Philosophy* 3 (1988), pp. 123-155.

Hutchins, Edwin, *Culture and Inference: A Trobriand Case Study* (Cambridge: Harvard University Press, 1980).

Jarvie, I. C. and J. Agassi, "The Problem of the Rationality of Magic," in *Rationality*, Brian Wilson, ed. (New York: Harper and Row, 1970), pp. 172-193.

____,"The Rationality of Dogmatism," in *Rationality Today*, T. Geraets, ed. (Ottowa: University of Ottowa Press, 1979), pp. 353-362.

Johnson-Laird, P. N., *Mental Models* (Cambridge: Cambridge University Press, 1983).

Johnson-Laird, P. N., P. Legrenzi, and M. S. Legrenzi, "Reasoning and a Sense of Reality," *British Journal of Psychology* 63 (1972), pp. 395-400.

Kahneman, Daniel, Paul Slovic and Amos Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases* (Cambridge: Cambridge University Press, 1982).

Kahneman, Daniel and Amos Tversky, "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology* 3 (1972). Reprinted in *Judgment under Uncertainty: Heuristics and Biases*, pp. 32-47.

Kimura, Mooto, "The Neutral Theory of Evolution," *Scientific American* 240:5 (1979), pp. 98-126.

Kitcher, Philip, *Vaulting Ambition: Sociobiology and the Quest for Human Nature* (Cambridge: MIT Press, 1985).

Kuhn, Thomas, "Logic of Discovery or Psychology of Research," in *The Essential Tension* (Chicago: University of Chicago Press, 1977), pp. 266-292.

_____, "Reflections on My Critics," in *Criticism and the Growth of Knowledge*, Imre Lakatos and Alan Musgrave, eds. (London: Cambridge University Press, 1970), pp. 231-278.

_____, *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1970).

Laudan, Larry, *Progress and Its Problems* (Berkeley: University of California Press, 1977).

Leslie, Alan, and Uta Frith, "Prospects for a Cognitive Neuropsychology of Autism: Hobson's Choice," *Psychological Review* 97 (1990), pp. 122-131.

Lewontin, Richard, "Caricature of Darwinism," *Nature* 266 (March 1977), pp. 283-284.

_____,"The Evolution of Cognition," in *Thinking: An Invitation to Cognitive Science*, volume three, Daniel Osherson and Edward E. Smith, eds. (Cambridge: MIT Press, 1990), pp. 229-246.

Lipton, Peter, and Nicholas Thompson, "Comparative Psychology and the Recursive Structure of Filter Explanations," *International Journal of Comparative Psychology* 1 (Summer 1988), pp. 215-244.

Lycan, William, *Judgment and Justification* (Cambridge: Cambridge University Press, 1988).

Lyons, William, *The Disappearance of Introspection* (Cambridge: MIT Press, 1986).

Mackie, J. L., "Self-Refutation—A Formal Analysis," *The Philosophical Quarterly* 14 (1964). Reprinted in *Logic and Knowledge: Selected Papers of J. L. Mackie*, Joan Mackie and Penelope Mackie, eds. (Oxford:Oxford University Press, 1985), pp. 54-67.

Macnamara, John, *A Border Dispute* (Cambridge: MIT Press, 1986).

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

Manktelow, K. I., and J. Evans, "Facilitation of Reasoning by Realism: Effect or Non-Effect," *British Journal of Psychology* 70 (1979), pp. 477-488.

Marr, David, *Vision: A Computational Investigation into Human Representation and Processing of Visual Information* (San Francisco: Freeman, 1982).

Millikan, Ruth, "Biosemantics," *Journal of Philosophy* 86 (1989), pp. 281-297.

_____, *Language, Thought and Other Biological Categories* (Cambridge: MIT Press, 1987).

_____, "Naturalist Reflections on Knowledge," *Pacific Philosophical Quarterly* 65 (1984), pp. 315-334.

Nisbett, Richard, and E. Borgida, "Attribution and the Psychology of Prediction," *Journal of Personal and Social Psychology* 32 (1975), pp. 932-943.

Nisbett, Richard, David Kranz, Christopher Jepson, and Ziva Kunda, "The Use of Statistics in Everyday Inductive Reasoning," *Psychological Review* 90 (1983), pp. 339-363.

Papineau, David, *Reality and Representation* (Oxford: Basil Blackwell, 1987).

Parfit, Derek, *Reasons and Persons* (Oxford: Oxford University Press, 1984).

Parker, S. T., and K. R. Gibson, *Language and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives* (Cambridge: Cambridge University Press: 1990).

Piattelli-Palmarini, Massimo, "Evolution, Selection and Cognition," *Cognition* 31 (1989), pp. 1-44.

Pinker, Steven, and Paul Bloom, "Natural Language and Natural Selection," *Behavioral and Brain Sciences* 13 (December 1990), pp. 707-784.

Pollard, Paul, "Natural Selection for the Selection Task: Limits to Social Exchange Theory," *Cognition* 36 (August 1990), pp. 195-204.

Popper, Karl, *Conjectures and Refutations* (NY: Basic Books, 1962).

_____, "Evolutionary Epistemology," in *Evolutionary Theory: Paths into the Future*, J.W. Pollard, ed. (London: Wiley and Sons, 1984).

_____, *The Logic of Scientific Discovery* (New York: Basic Book, 1959).

_____, "Natural Selection and the Emergence of Mind," *Dialectica* 32 (1978), pp. 339-355. Reprinted in *Evolution, Theory of Rationality and the Sociology of Knowledge*, G. Radnitsky and W.W. Bartley III, eds., pp. 139-156.

_____, *Objective Knowledge: An Evolutionary Approach* (Oxford: Oxford University Press, 1972).

Putnam, Hilary, "The 'Innateness Hypothesis' and Explanatory Models in Linguistics," *Synthese* 17 (1967), pp. 12-22. Reprinted in *Readings in the Philosophy of Psychology*, Ned Block, ed,. volume two (Cambridge: Harvard University Press, 1981), pp. 292-299.

_____, "The Meaning of 'Meaning,'" in *Mind, Language and Reality: Philosophical Papers*, volume 2 (Cambridge, England: Cambridge University Press, 1975), pp. 215-271.

Quine, W. V. O., *Ontological Relativity and Other Essays* (NY: Columbia University Press, 1969).

_____, "Two Dogmas of Empiricism," in *From a Logical Point of View* (Cambridge: Harvard University Press, 1961), pp. 20-46.

_____, *Word and Object* (Cambridge: MIT Press, 1960).

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

Rachlin, Howard, A. W. Logue, John Gibbon, and Marvin Frankel, "Cognition and Behavior in Studies of Choice, *Psychological Review* 93 (1986), pp. 33-45.

Rawls, John, "The Independence of Moral Theory," *Proceedings and Addresses of the American Philosophical Association* 48 (1974-1975), pp. 5-22.

_____, *A Theory of Justice* (Cambridge: Harvard University Press, 1971).

Ruse, Michael, *Taking Darwin Seriously* (Oxford: Basil Blackwell, 1986).

Seligman, Martin, and Joane Hager, *The Biological Boundaries of Learning* (NY: Appleton-Century Crofts, 1972).

Skagestad, Peter, "Taking Evolution Seriously: Critical Comments on D.T. Campbell's Evolutionary Epistemology," *Monist* 61 (October 1978), pp. 611-621.

Skyrms, Brian, *Choice and Chance* (Belmont, CA: Wadsworth, 1986).

Slovic, Paul, Baruch Fischhoff, and Sarah Lichtenstein, "Cognitive Processes and Societal Risk Taking," in *Cognition and Social Behavior*, J. S. Carroll and J. W. Payne, eds. (Hillsdale, NJ: Ersbaum, 1976), pp. 165-184.

Sober, Elliott, "The Evolution of Rationality," *Synthese* 46 (1981), pp. 95-120.

_____, *The Nature of Selection* (Cambridge: MIT Press, 1984).

_____, "Psychologism," *Journal of Social Behavior* 8 (1978), pp. 165-191.

Stein, Edward, "Getting Closer to the Truth: Realism and the Metaphysical and Epistemological Ramifications of Evolutionary Epistemology," in *Evolution, Cognition and Realism*, Nicholas Rescher, ed., pp. 119-129.

Stein, Edward, and Peter Lipton, "Where Guesses Come From: Evolutionary Epistemology and the Anomaly of Guided Variation," *Biology and Philosophy* 4 (1990), pp. 33-56.

Stent, Gunther, "You Can Take the Ethics Out of Altruism but You Can't Take the Altruism Out of Ethics," *Hastings Center Report* 7 (1977), pp. 33-36.

Stich, Stephen, "Could Man Be an Irrational Animal?," *Synthese* 64 (1985), pp. 115-135. Reprinted in *Naturalizing Epistemology*, Hilary Kornblith, ed. (Cambridge: MIT Press, 1985), pp. 249-267.

_____, "Dennett on Intentional Systems," *Philosophical Topics* 12 (1981), pp. 39-62.

_____, "Empiricism, Innateness, and Linguistic Universals," *Philosophical Studies* 33 (1978), pp. 273-286.

_____, *Fragmentation of Reason* (Cambridge: MIT Press, 1990).

_____, "Relativism, Rationality and the Limits of Intentional Description," *Pacific Philosophical Quarterly* 65 (1984), pp. 211-235.

Stich, Stephen, and Richard Nisbett, "Justification and the Psychology of Human Reasoning," *Philosophy of Science* 47 (1980), pp. 188-202.

Sternberg, Saul, "High-speed Scanning in Human Memory," *Science* 153 (1966), pp. 652-654.

Templeton, A., "Adaptation and the Integration of Evolutionary Forces," in *Perspectives on Evolution*, R. Milkman, ed. (Sunderland, MA: Sinauer, 1982).

Thagard, Paul, "Against Evolutionary Epistemology," in P. D. Asquith and R. N. Giere, eds., *PSA 1980* (1980) pp. 187-196.

Thagard, Paul, and Richard Nisbett, "Rationality and Charity," *Philosophy of Science* 50 (1983), pp. 250-267.

Tooby, John, "The Emergence of Evolutionary Psychology," in *Emerging Synthesis in Science*, D. Pines, ed. (Redwood City, CA: Addison-Wesley, 1988), pp. 67-75.

Tooby, John, and Leda Cosmides, "Evolutionary Psychology and the Generation of Culture: Part I: Theoretical Considerations," *Ethology and Sociobiology* 10 (1987), pp. 29-49.

Tversky, Amos, "Features of Similarity," *Psychology Review* 84 (July 1977), pp. 327-352.

Tversky, Amos and Daniel Kahneman, "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment," *Psychology Review* 90 (October 1983), pp. 293-315.

_____, "Judgment under Uncertainty: Heuristics and Biases," *Science* 85 (1974), pp. 1124-1131. Reprinted in *Judgment under Uncertainty: Heuristics and Biases*, Daniel Kahneman, Paul Slovic and Amos Tversky, ed., pp. 3-20.

_____, "On the Interpretation of Intuitive Probability: A Reply to Jonathan Cohen," *Cognition* 7 (1979), pp. 409-411.

Walker, Steven, *Animal Thought* (Boston:Routledge, 1983).

Wason, P. C., "Reasoning," in *New Horizons in Psychology*, B. Foss, ed. (Middlesex, England: Penguin, 1966), pp. 135-151.

_____, "Reasoning about a Rule," *Quarterly Journal of Experimental Psychology* 20 (1968), pp. 273-281.

_____, "Regression in Reasoning?," *British Journal of Psychology* 60 (1969), pp. 471-480.

Wason, P. C., and P. N. Johnson-Laird, "A Conflict Between Selecting and Evaluating Information in an Inferential Task," *British Journal of Psychology* 61 (1970), pp. 509-515.

_____, *Psychology of Reasoning: Structure and Content* (Cambridge: Harvard University Press, 1972).

Wason, P. C., and D. Shapiro, "Natural and Contrived Experience in a Reasoning Problem," *Quarterly Journal of Experimental Psychology* 23 (1971), pp. 63-71.

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

Williams, George C., *Adaptation and Natural Selection* (Princeton: Princeton University Press, 1966).

Wilson, Edward O., *Sociobiology: The New Synthesis* (Cambridge: Harvard University Press, 1975).

Wilson, Edward O., and Charles Lumsden, *Genes, Minds and Culture* (Cambridge: Harvard University Press, 1981).

_____, *Promethean Fire* (Cambridge, Harvard University Press, 1983).

Wilson, N. L., "Substances Without Substrata," *Review of Metaphysics* 12 (1959), pp. 521-539.