

## Chapter One:

### Introduction

Suppose a friend of yours is talking about his new friend Linda. He tells you that Linda is single, thirty-one, bright, outspoken and that, as a college student, she majored in philosophy and was concerned with issues of social justice. Given all your friend has told you about her, rank the following statements in order of the likelihood that they will be true of Linda:

- (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.

If you are like most people, you ranked (3) as being more probable than (2). This, however, is a mistake. (3) cannot be more probable than (2), because (2) is true whenever (3) is. Although this fact seems straightforward once I have pointed it out, people seem to systematically avoid it.<sup>1</sup> In so doing, people are making a mistake in reasoning. According to experiments done over the past few decades, humans make similarly significant errors in various realms of reasoning: logical reasoning, probabilistic reasoning, similarity judgments, and risk-assessment to name a few.<sup>2</sup> Together these experiments are taken to show that humans are irrational.

---

<sup>1</sup> This example is adapted from Amos Tversky and Daniel Kahneman, 'Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment', *Psychological Review* 90 (October 1983), 293-315. For further discussion of this experiment, see Chapter Three, section II below.

<sup>2</sup> A representative cross-section of the reasoning experiments can be found in Daniel Kahneman, Paul Slovic and Amos Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases* (Cambridge: Cambridge University Press, 1982).

The observation that humans are irrational is perhaps more commonplace than that humans have two legs, even though the latter seems more obvious. Even in the face of evidence from experiments like the *conjunction experiment* sketched above, many want to resist the thesis that humans are irrational. Some philosophers and psychologists have developed creative and appealing arguments that these experiments are mistaken or misinterpreted because humans *must* be rational. Any one of these arguments, if successful would provide important insight into human nature. These arguments would also entail that there are limits to what science can show, namely science cannot show humans to be irrational. Finally, these arguments for human rationality have significant implications for epistemology, philosophy of science, philosophy of mind and philosophy of language.

In what follows, I examine the various arguments for human rationality and for the existence of limits to cognitive science and science in general. I attempt to show that these arguments fail; cognitive science can and should play a role in determining whether or not humans are rational. My discussion has implications for the distinction between empirical and conceptual knowledge, the proper relationship between philosophy and science, and for the project of epistemology. In particular, I will suggest that the traditional approach to theory of knowledge errs by ignoring the important role science plays in epistemology.

## I. What Is It to Be Rational?

### A. Three Ways to Be Rational

There are two truisms that seem in tension but, at the same time, seem to happily coexist in our commonsense view of humans and rationality. On the one hand, we agree with Aristotle that man is a rational animal, while on the other hand, we agree with Freud that humans are irrational. This apparent tension can be resolved by distinguishing among several different senses of "rational". First, when talking about rationality, one might

simply be referring to reasoning ability; humans are rational in the sense that we consciously and explicitly reason, we give arguments for the things we believe, and so on. Perhaps this is what Aristotle meant when he said that man is a rational animal. Second, one might use "rational" to denote *perfect* reasoning; humans are thus *irrational* in the sense that we frequently make mistakes in reasoning due, more often than not, to the vicissitudes of the human condition. For example, we make mistakes in reasoning because our behavior is influenced by repressed sexual desires. Perhaps this is what is behind the Freudian truism that humans are irrational. Glossed in these ways, the two truisms are perfectly compatible. It is consistent with the fact that humans reason that sometimes humans make mistakes in reasoning; in fact, the truth of the second fact depends on the truth of the first.

There is a third sense of "rational" that is used when talking about humans. This sense of rational allows that humans can make reasoning errors, but attributes these mistakes to forces that interfere with human reasoning rather than to mistakes internal to our reasoning process. There is a quite intuitive distinction at work here. To simplify matters, consider the following observation: there are two different explanations behind a person's inability to, say, answer a question. Typically, I have no trouble telling you precisely how old I am, but, on certain occasions, say when I am quite tired, under the influence of mind-altering substances, or preoccupied, I may have difficulty reporting my age, that is, I may hesitate, blurt out the wrong answer, or the like. In contrast, if you ask a total stranger how old I am, she will not know the answer; at best, she could guess, but, if she has never laid eyes on me and knows next to nothing about me, her guess will be a wild one. The stranger lacks knowledge of my age, while I have the knowledge but, under certain circumstances, I am unable to deliver it because of interfering factors. In terms that I will explain at some length below, when she gets my age wrong, she is making a *competence* error but when I get my age wrong, I am making a *performance* error. The third sense of rationality appeals to this distinction. According to this sense of

rationality, humans are rational if the only mistakes in reasoning we make are attributable to interferences with human reasoning rather than to human reasoning ability itself. This sense of rationality is particularly interesting and it is this sense that philosophers and psychologists have in mind when they debate whether or not humans are rational, and in particular, it is this sense of rational that is involved when psychological experiments (like the one about Linda sketched above) are cited as 'hav[ing] bleak implications for human rationality.'<sup>3</sup>

### B. The Standard Picture of Rationality

I call the claim that humans are rational *the rationality thesis* and the claim that humans are irrational *the irrationality thesis*. Both of these claims are typically based on what I call *the standard picture of rationality*. According to this picture, to be rational is to reason in accordance with principles of reasoning that are based on rules of logic, probability theory and the like. If the standard picture of reasoning is right, principles of

---

<sup>3</sup> Richard Nisbett and Eugene Borgida, 'Attribution and the Psychology of Prediction', *Journal of Personal and Social Psychology* 32 (1975), 935. Others who have claimed that such experiments support the conclusion that humans are irrational include: Daniel Kahneman and Amos Tversky, 'Subjective Probability: A Judgment of Representativeness', *Cognitive Psychology* 3 (1972), reprinted in *Judgment under Uncertainty*, 46 (in *Judgment* anthology), '[F]or anyone who would wish to view man as a reasonable intuitive statistician, such results are discouraging'; Paul Slovic, Baruch Fischhoff and Sarah Lichtenstein, 'Cognitive Processes and Societal Risk Taking', in *Cognition and Social Behavior*, John Carroll and John Payne, eds. (Hillsdale, NJ: Erlbaum, 1976), 173-4, '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), 340, 'people commit serious errors of inference'.

reasoning that are based on such rules are *normative principles of reasoning*, namely they are the principles we *ought* to reason in accordance with.

The standard picture of rationality is not the only possible picture of rationality. In Chapter Seven, I will consider others. As described, however, the standard picture of rationality is intuitively very plausible. It seems certain that the principles which we think are the normative principles of reasoning are in fact the normative principles of reasoning. To appreciate the appeal of the standard picture, compare the principles of reasoning with mathematics. Consider the statement 'Two plus two equals four.' This statement seems intuitively plausible. A theory of mathematics that says two plus two does not equal four would be highly implausible. Principles of reasoning based on rules of logic, probability and the like seem equally well-established.

Assuming that the standard picture of rationality is the right picture, we might wonder why it is. There are various theories, some of which I will return to in subsequent chapters. For now, consider what makes mathematical facts true. According to Platonism, which is perhaps the standard view of mathematics, mathematical facts are true independent of cognitive operations and independent of any relation to human minds. Two plus two equals four regardless of whether humans believe it or not and even before humans existed. Although Platonism has been criticized on various grounds, there is at least the quite strong appearance that it is true. A non-Platonist theory must at least explain the strong intuitive appeal of Platonism. Platonism seems equally strong with respect to the rules of logic. The mathematical statement 'Two plus two equals four.' seems of the same status as the logical statement "'Bill Clinton and Hilary Clinton live in the White House" entails "Bill Clinton lives in the White House.'" The rule of mathematics that says multiplication of  $x$  by  $y$  is equivalent to the sum of  $y$   $x$ 's is of the same status as the following rule of logic:

**and-elimination rule:** the truth of the conjunction of two statements entails the truth of each statement in the conjunction.

The standard picture of rationality does not require the truth of Platonism; for example, the standard picture, as described, is compatible with the view that the principles of rationality are whatever we think they are. The standard picture of rationality does, however, fit nicely with Platonism's intuitive appeal when it is combined with the view that we are equipped to access mathematical facts, logical truths and the normative principles of reasoning. The principles that we think are the normative principles of reasoning match the principles that are the normative principles of reasoning because we have the capacity for determining what the normative principles are. For now, I will assume that the standard picture of rationality, or an account roughly like it, is true. As I proceed, however, I will raise several problems for this view.

Note that the principles of logic do not, on their own, provide an account of how we ought to reason. Reasoning involves beliefs and the rules of logic say nothing about beliefs. According to the standard picture of rationality, the principles of reasoning that we ought to follow are based on rules of logic. Consider the statement

(a) Bill Clinton is the president and Al Gore is the vice-president.

This statement entails both the statement

(b) Bill Clinton is the president.

and the statement

(c) Al Gore is the vice-president.

This example is an instance of the and-elimination rule of logic. The and-elimination rule gives rise to the following principle of reasoning:

**and-elimination principle:** if you believe the conjunctive statement **A and B**,

you should believe both the statement **A** and the statement **B**.

According to this principle, if you believe the statement (a) then you should also believe the statement (b) and the statement (c). Note that I have distinguished between rules of logic and principles of reasoning that stem from them. Rules of logic apply to statements and determine the logical relations among them; principles of reasoning that stem from

rules of logic apply to beliefs and determine the relations among them. Some, but certainly not all, principles of reasoning are based on rules of logic. According to the standard picture of rationality, principles of reasoning based on rules of logic are normative principles of reasoning.

As another example, consider the following rule of logic:

**modus ponens:** A and **if A, then B** together entail **B**.

This gives rise to the following normative principle of reasoning:

**modus ponens principle:** if you believe **A** and you believe **if A, then B**, you should believe **B**.

So, for example, if you believe

If unemployment rate goes up, then the stock market will fall.

and you believe

The unemployment rate will go up.

then you should also believe

The stock market will fall.

Not all principles of reasoning are based on rules of logic; some normative principles of reasoning are based on the rules of probability theory. Consider the following rule of probability:

**conjunction rule:** the probability of some event **A** occurring cannot be less than the probability of **A** and some other event **B** both occurring.

This rule of probability gives rise to the following normative principle of reasoning:

**conjunction principle:** you should not attach a lesser degree of probability to an event **A** than you do to both the event **A** and the (distinct) event **B** occurring.

For example, the conjunction principle says that if you believe

There is a fifty percent chance that, tomorrow, the temperature will be forty degrees and it will rain.

then you should *not* believe

There is a *less* than fifty percent chance that it will rain tomorrow.

It was the conjunction principle that was violated by subjects in the experiment involving Linda. It cannot be more likely that Linda is a bank teller and a feminist than it is that she is a bank teller because in every case in which Linda is a bank teller and a feminist she is also a bank teller.

So far, I have just enumerated a couple of principles of reasoning that are plausibly normative and seemingly important. It would be nice, however, if I could give a complete list of the normative principles of reasoning. Enumerating the normative principles of reasoning is a notoriously difficult task that goes back to the early days of epistemology. Producing a complete list of the norms of reasoning would go quite far towards developing a complete and true theory of knowledge (the remaining task—itsself a rather difficult one—would be to explain why these principles are the norms and not some others—that is, to *justify* these principles). Developing a complete list of the norms of reasoning is, however, beyond the scope of this project. Given the difficulty of this task, it is fortunate that I do not need to have a complete list of the norms of reasoning in order to determine what kind of question it is to ask whether or not humans reason according to the norms or to determine the proper relationship between epistemology and cognitive science, and, more generally, epistemology and science. Instead, it will suffice to enumerate a couple of principles of reasoning that are plausibly normative and seemingly important.

### C. Reasoning in Accordance with the Normative Principles

Having sketched what the normative principles of reasoning are, the next obvious question concerns what it is to reason in accordance with these norms. A tempting answer to this question is that to reason in accordance with the norms of reasoning is to *always* do what the norms would dictate, for example, to always believe **B** when you believe **A** and **if A, then B**, to always believe that the probability of **A** is greater than or

equal to the probability of **A and B**, and so on. This quick answer cannot be what friends of the rationality thesis have in mind; always doing what the norms of reasoning dictate is clearly not a necessary condition for counting someone as rational since 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 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 I know George is in his office only on Fridays, and then expect me to admit my mistake and confess that it must be Friday. Assuming I did this, when you look 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 continue to deny that today is Friday and do 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. This suggests that more is involved in failing to be rational than just failing to always do what the norms dictate; sometimes failing to behave in a way that matches the norms is not a sign of irrationality, but rather is just due to having made a mistake (or, in terms that I will explain below, sometimes failing to behave in a way that matches the norms is a performance error rather than being indicative of having an irrational reasoning competence). It seems that to reason in accordance with a rule of reasoning is to reason in such a way that your behavior is typically best explained by your following this rule even if you sometimes make mistakes.

Behind the notion of a mere mistake is the idea that making a mistake involves a *momentary lapse*, a divergence from some typical behavior. This is in contrast to

attributing a divergence from a norm to reasoning in accordance with principles that diverge from the normative principles of reasoning. Behavior due to irrationality connotes a *systematic* divergence from the norm. It is this distinction between mere mistakes and systematic violations (between performance errors and competence errors) that is implicitly assumed when the reasoning experiments are cited as evidence for the irrationality thesis. It is also implicitly assumed by friends of the rationality thesis when they *deny* that the reasoning experiments are relevant to whether or not the rationality thesis is true.

Consider, by way of example, the conjunction experiment. Subjects in this experiment are given a description of Linda and asked 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 principle dictates that my assessment of the probability that Linda is a bank teller should be greater than or equal to my assessment of the probability that Linda is both a bank teller and a feminist. The conjunction experiment shows that the conjunction rule is frequently violated by individuals and across the human population in general. Friends of the irrationality thesis say that the best explanation of these results is that humans follow some principle other than the conjunction principle; interpreted in this way, the results count in favor of the irrationality thesis. This explanation supports the picture of humans as making systemic errors in reasoning rather than making mere mistakes: we do not generally reason in accordance with the conjunction rule but sometimes make mistakes; rather, we systematically violate the conjunction rule because we reason in accordance with some other principle of reasoning that is not a normative principle of reasoning.

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 an even better way of describing what is at issue between the rationality thesis and the irrationality thesis that involves borrowing the linguist's distinction between competence and performance.<sup>4</sup>

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 or write ungrammatical 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, an interference that prevents a person from engaging in linguistic behavior that is in accordance with linguistic competence. These interferences involve 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 say that *all* divergences from the norms of reasoning are performance errors and, as such, these divergences are not indicative of an underlying ability to reason. Defenders of the irrationality thesis agree that the competence-performance distinction is applicable to the realm of reasoning, but they

---

<sup>4</sup> 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).

deny that our *reasoning competence*<sup>5</sup> matches the norms of reasoning; they offer alternative accounts of human reasoning competence, accounts according to which we are not rational. For example, friends of the irrationality thesis would claim that the conjunction experiment shows that following the conjunction rule is *not* part of our reasoning competence; in general, they would argue that our reasoning competence does not match the norms of reasoning and that thereby humans are irrational. The notion of competence is crucial to getting the debate between the rationality thesis and the irrationality thesis off the ground; it is especially crucial to reconciling the rationality thesis with the results of the reasoning experiments. I will explore the analogy between linguistic competence and reasoning competence at length in Chapter Two. Chapter Two begins with an extensive discussion of linguistic competence. Linguistic competence is a relatively well-developed notion; it can be usefully compared and contrasted with reasoning competence. This comparison is the topic of the remainder of Chapter Two. I argue that there are some important similarities between linguistic competence and reasoning competence, but there is one important difference—linguistic norms (that is, principles of grammaticality) are clearly indexed to linguistic competence while

---

<sup>5</sup> John Macnamara, *A Border Dispute* (Cambridge: MIT Press, 1986) uses the term "mental logic", Stephen Stich, *Fragmentation of Reason* (Cambridge: MIT Press, 1990), uses the term "psycho-logic", and L. Jonathan Cohen, 'Can Human Irrationality Be Experimentally Demonstrated?', *Behavioral and Brain Sciences* 4 (1981), 317-70, uses the term "cognitive competence", for roughly the same concept that I prefer to call "reasoning competence". I opt for "reasoning competence" because it emphasizes the importance of the notion of competence as borrowed from linguistics but does not suggest as broad a notion as Cohen's term "cognitive competence". "Cognitive competence" seems like it would include linguistic competence while "reasoning competence" does not. For my purposes, I want to avoid suggesting that linguistic competence is part of our underlying ability to reason while at the same time suggesting an analogy with linguistic competence.

principles of reasoning are not obviously indexed to reasoning competence. For now, I adopt the following terminology for discussing the two theses about human rationality: the rationality thesis says that human reasoning competence matches the normative principles of reasoning (that is, the rules embodied in our reasoning competence are the same as those that we ought to follow), while the irrationality thesis says that human reasoning competence diverges from the norms (that is, the rules embodied in our reasoning competence are different from those we ought to follow).

#### D. Whose Rationality?

Thus far, I have been talking about human rationality as if it is clear whose reasoning abilities I am talking about. One might reasonably ask whether the rationality thesis requires that *all* humans reason in accordance with the normative principles of reasoning, that *most* humans reason in accordance with the normative principles, that *more than half* of us do so, or just that *one* person reasons in accordance with the norms. By talking about human rationality in terms of human reasoning competence, the rationality and irrationality theses concern the reasoning *capacities* of all *normal* humans. The rationality thesis says that all normal humans have the capacity to reason in accordance with the normative principles of reasoning. I am assuming that normal humans have basically the same type of reasoning competence in the same way that normal humans typically have the same digestive system or visual capacity.

Given this, there are various ways in which the rationality thesis is compatible with an individual human not reasoning in accordance with the norms. First, she might not be a normal human in term of reasoning capacity. One can talk about human reasoning capacity while allowing that there may be people with impaired reasoning capacities in much the same way as one can talk about human vision and human visual capacity while allowing that there are blind, color blind, and other visually-impaired people. This means that the rationality thesis is compatible with there being some humans who are not normal

because they have a reasoning competence that does not match the normative principles of reasoning. Second, a person might have a certain capacity for reasoning but fail to make use of this capacity. A normal child who is not exposed to any language by puberty will not develop a full-fledged natural human language,<sup>6</sup> in spite of the fact that, as a normal human, she has the capacity for language. Similarly, it is compatible with the rationality thesis that a normal person might not reason in accordance with the normative principles if she has the capacity for reasoning but does not make use of it. Human reasoning competence might, for example, be like human linguistic competence in that certain environmental inputs are required for the competence to develop to its capacity. The rationality thesis is compatible with there being a normal human who does not reason in accordance with the norms because she failed to get the appropriate environmental inputs. The rationality thesis, thus clarified, says that all normal humans have a reasoning competence that gives us the capacity to reason in accordance with the normative principles of reasoning, even though some normal humans may not attain their full capacity.

Conversely, the irrationality thesis says that all normal humans have a reasoning competence such that we do not have the capacity to reason in accordance with the normative principles. The irrationality thesis is compatible, however, with the claim that *sometimes* people reason in the way that they would if they were following the normative principles of reasoning. Consider the following non-normative principle of reasoning:

**asymmetrical and-elimination principle:** if you believe the conjunctive statement **A and B**, you should believe **A** but *not B*.

A person who has this principle in her reasoning competence would infer

(b) Bill Clinton is the president.

---

<sup>6</sup> Steven Pinker, *The Language Instinct: How the Mind Creates Language* (New York: Morrow, 1994),

from the statement

(a) Bill Clinton is the president and Al Gore is the vice-president.

In so doing, she would be reasoning in the same way that she would if she were following the and-elimination principle. From (a), she would also infer

(d) Al Gore is not the vice-president.

In so doing, she would not be reasoning in the same way that she would if she were following the and-elimination principle. This explains how the conjunction experiment (the one involving Linda) can be taken to support the irrationality thesis even though some subjects in the experiment correctly say that Linda is more likely to be a bank teller than she is to be a bank teller and a feminist.

What if only one person (or just a few people) always follows the normative principles of reasoning? Would this be enough to establish the truth of the rationality thesis? This depends on the details of the situation. There are three ways that only one person might reason in accordance with the normative principles of reasoning. First, every human might have the capacity to reason in accordance with the normative principles of reasoning, but only one person might fulfill this capacity. If this were the case, on my understanding of rationality, humans would be rational. Second, every human might not have the normative principles of reasoning in their reasoning competence but one person might simply be lucky in that her reasoning behavior matches the reasoning behavior of a person who has the normative principles of reasoning in her reasoning competence. In this case, humans would be irrational. Third, all humans but one might lack the normative principles of reasoning in their reasoning competence. In this case, the person with the normative principles in her reasoning competence would be non-normal in the way that a person without a stomach or who cannot see is non-normal. In this case, humans would be irrational.

## E. Why Is This Interesting?

Whether or not humans are rational is an interesting and important question about human nature. My discussion in subsequent chapters will emphasize the connections between human rationality and epistemology, philosophical psychology, and philosophy of language, on the one hand, and between human rationality and cognitive science, evolutionary theory, neuroscience and theory of computation on the other. Human rationality is interesting and important for many other reasons as well. Aristotle thought that rationality is part of our human essence and a feature of human excellence. Is he right on either score? Some political theorists have argued that the viability of democracy depends of the rationality of humans as political agents. If humans are irrational, is democracy a bad form of government? Economics, on the most widely accepted view of the field, requires the assumption that humans are rational. If humans are irrational, is all of economic theory undermined? In this book, I will not attempt to develop the connections between my project and these other issues. I will not even attempt to offer a defense of the relevance of my project to these issues. I mention them here to point to various reasons outside of epistemology, philosophy of science and philosophy of psychology as to why we should care whether humans are rational.

I should also mention that the sort of irrationality that I am interested in this book is at least a bit different from another kind of irrationality, what I call *irrational action*, that is, action that seems to go against an agent's goals, best interests, and the like.<sup>7</sup> An example of an irrational action is continuing to smoke even though you want to live as long as

---

<sup>7</sup> Irrational action has been the subject of numerous recent philosophical inquiries, for example, David Pears, *Motivated Irrationality* (Oxford: Oxford University Press, 1984); Alfred Mele, *Irrationality: An Essay in Akrasia, Self-Deception and Self-Control* (Oxford: Oxford University Press, 1987); and Brian McLaughlin and Amélie Rorty, eds., *Perspectives in Self-Deception* (Berkeley: University of California Press, 1988).

possible and you know that smoking decreases life expectancy. You might continue to smoke without violating a normative principle of reasoning; your irrational action might be attributable to a physical addiction to nicotine, a weakness of the will, or something other than a deviation from a norm. Irrationality in the sense of irrational action is my concern in this book insofar as it relates to irrational reasoning.

## II. Traditional, Naturalized and Descriptive Epistemology

Epistemology is the study of knowledge: what it is, how it is attained, and who has it.

Typically, knowledge is understood to be a species of belief; if I *know* Bill is in Washington, then I must *believe* Bill is in Washington. The reverse is not true; each of us has many beliefs that do not count as knowledge. One way you can have a belief that is not knowledge is if you believe a statement that is false. If I *believe* that the moon is made of green cheese, I do not *know* the moon is made of green cheese (even if I feel quite confident about my belief), because my belief is false, the moon is not made of green cheese. Another way that I can have a belief that is not knowledge is if I believe something that I am not justified in believing. Suppose you ask me what the first name of the prime minister of England is. Thinking that "John" is a common name for Englishmen and that most prime ministers are men, I form the belief that "John" is the prime minister's first name; accordingly, I can be said to believe that the prime minister's first name is John. As John Major is the prime minister, my belief would be true. My belief would not however be *justified* (roughly because my belief is not appropriately connected to the fact that John Major is the prime minister) and hence would not be an instance of knowledge. At a minimum, then, to know something is to have a true belief that is justified.<sup>8</sup>

---

<sup>8</sup> Edmund Gettier, 'Is Justified True Belief Knowledge?', *Analysis* 23 (1963), 121-3, argues that a belief can be true and justified but still not an instance of knowledge. A veritable cottage industry in philosophy

In order to gain knowledge, one has to have beliefs. Beliefs can be acquired in various ways including directly through the senses (I believe that it is raining outside because I see rain falling when I look outside), through testimony (I believe that it is raining outside because a trusted friend tells me that it is), and through reasoning (I believe that it is raining because I see that my mother has taken out her umbrella, and I believe that she only takes out her umbrella when it is raining). Because reasoning is an important way that we manipulate beliefs and acquire knowledge, the study of human rationality is an important part of epistemology.

Rene Descartes started his meditations on human knowledge by noting that, in the past, he believed many things that later proved false.<sup>9</sup> Given this fact, he endeavored to reflect on various principles for the acquisition of beliefs, rejecting those that have led him astray in the past or that might lead him astray in the future and embracing only those principles that will lead him to the truth. In order to do this, Descartes looked for a firm and immobile point, a truth or truths on which he could ground principles of belief acquisition and maintenance that would insure that all his beliefs are knowledge. Descartes' project was to develop an account of how we should arrive at our beliefs. He tried to do so independently of experience and without looking at the world because experiential access to the world is fallible and thus not firm ground on which to develop

---

developed in response to Gettier's article to attempt to fill in the additional condition(s) for knowledge in addition to truth and justifiedness. Some of the discussions of this article, as well as the article itself, are reprinted in Paul Moser, ed., *Empirical Knowledge: Readings in Contemporary Epistemology* (Savage, MD: Rowman and Littlefield, 1986), Part II, 231-70. Addressing Gettier's challenge to the task of defining knowledge is beyond the scope of the present project. I mention it here to indicate that it is not completely clear what the criterion for knowledge is.

<sup>9</sup> Rene Descartes, *Meditations Concerning First Philosophy*.

an account of how we should arrive at our beliefs. The Cartesian project is central to what I will call *traditional epistemology*.

Some, most notably Willard V. O. Quine,<sup>10</sup> think that the normative project of traditional epistemology should be rejected because it is hopeless. They say that there is no way to ground our beliefs on a firm and immobile point. If this is right, the traditional project of epistemology might be replaced by the project of describing how we come to believe things. Traditional epistemology would, on this view, be replaced by 'the science of belief', what might be called *descriptive epistemology*. The project is *descriptive* rather than *normative* in that it tells us how we proceed with respect to beliefs but *not* how we *should* proceed. Normative epistemology, a part of philosophy as traditionally construed is to be replaced by descriptive epistemology, a part of science that is connected to psychology, biology and neuroscience.

What I call descriptive epistemology is often referred to as 'naturalized epistemology.' There is, however, another project for which I want to reserve that term. This project preserves a normative component for epistemology (like traditional epistemology) but at the same time it draws from science (like descriptive epistemology). *Naturalized epistemology* uses scientific evidence as part of the process of determining how we ought to acquire beliefs. Perhaps the best way to illustrate this project is Otto Neurath's metaphor of the ship. Imagine that we are on a ship that has some rotten planks. We want to repair the ship's planks. The best way to do this would be to bring the ship into dock, disembark, and, standing on firm ground, repair the ship's planks. But suppose that we are far from any dock. We can still repair the ship by standing on some planks while repairing the others. This project is not guaranteed to succeed because the planks that we

---

<sup>10</sup> Willard V. O. Quine, 'Epistemology Naturalized', in *Ontological Relativity and Other Essays* (NY: Columbia University Press, 1969), 69-90; reprinted in *Naturalizing Epistemology*, second edition, Hilary Kornblith, ed. (Cambridge: MIT Press, 1994), 15-32.

choose to stand on while making repairs on some others may themselves be rotten. Still, if we carefully chose the planks we stand on, we can be somewhat hopeful of success. Our beliefs are like the planks on the ship. Just as some of the ship's planks are rotten, some of our beliefs are false. We want to rid ourselves of false beliefs in the same way that we want to rid the ship of rotten planks. Traditional epistemology says that we can get rid of false beliefs by finding firm ground on which to base our beliefs. The arguments for descriptive epistemology say that the human epistemological condition is like a ship permanently at sea, there is no firm ground from which we can repair our epistemological ship. Naturalized epistemology says that, just as we can repair a ship at sea, we can get rid of some false beliefs by assuming the truth of some of our beliefs, typically the truth of some of our scientific beliefs.<sup>11</sup>

---

<sup>11</sup> Others have drawn this distinction in different ways and using different terms. Jaegwon Kim, 'What Is "Naturalized Epistemology?,"' in *Philosophical Perspectives*, volume 2, *Epistemology*, James Tomberlin, ed. (Atascadero, CA: Ridgeview, 1988); reprinted in *Naturalizing Epistemology*, 33-55, uses the phrase "Quine's naturalized epistemology" for the theory of knowledge that I call "descriptive epistemology" and the term "naturalism" for what I call "naturalized epistemology." Hilary Kornblith, 'Introduction: What is Naturalistic Epistemology?', in *Naturalizing Epistemology*, 1-14, uses "naturalized epistemology" for the conjunction of what I call "descriptive epistemology" and what I call "naturalized epistemology". What Kornblith calls "the strong replacement thesis" would be a thesis entailed by descriptive epistemology and what he calls "the weak replacement thesis", "psychologism", and "ballpark psychologism" would, in my terminology, be versions of naturalized epistemology. Philip Kitcher, 'The Naturalists Return', *Philosophical Review* 101 (1992), 53-114, uses the phrase "radical naturalism" for what I call "descriptive epistemology" and the phrase "conservative naturalism" for what I call "naturalized epistemology". Two problems with my terminology should be indicated. First, some people would deny that what I call "descriptive epistemology" is epistemology at all. For the moment, at least, I am open to the possibility that this is a misnomer. Second, my terminology makes the title of Quine's

Like traditional epistemology, naturalized epistemology has a normative component—it tells us which of our beliefs count as knowledge (in terms of the ship, it tells us which planks are rotten)—but, like descriptive epistemology, it has at least a partially empirical character. The empirical contribution to naturalized epistemology can come in various forms. For example, a naturalized epistemology, rather than explaining the justification of beliefs in terms of the logical properties of beliefs or the logical relations of contents of beliefs (which is what traditional epistemology does) might explain the justification of beliefs in terms of causal or law-like connections among psychological states or processes. The naturalized epistemologist can do this without requiring that justification must be spelled out in naturalistic terms.<sup>12</sup> Or, the naturalized epistemologist can see empirical facts as constraining or providing a useful way of testing epistemological theories formulated in non-empirical terms (that is, in terms of the logical properties of beliefs or the logical relations of contents of beliefs).<sup>13</sup>

The nature of the question 'Are humans rational?' is related in interesting ways to these three different ways (traditional, descriptive, and naturalized) of doing epistemology. Rationality is usually seen as a normative notion; to say that someone is rational is roughly to say that she reasons in the way she ought to. The traditional epistemologist agrees with this traditional picture of rationality as normative. Because reasoning is one of the ways we manipulate beliefs and because the traditional epistemologist thinks that there are particular ways that beliefs ought and ought not to be manipulated, the traditional epistemologist sees rationality as normative. He would add

---

seminal article 'Epistemology Naturalized' a misleading one; according to my terminology, 'Epistemology Descriptivized' would have been preferable (though less felicitous). I take it that neither of these problems are particularly serious.

<sup>12</sup> Kim, 'What Is "Naturalized Epistemology?"', and Kornblith, 'The Naturalists Return'.

<sup>13</sup> Kornblith, 'Introduction', 10-12.

that determining what is rational involves reflection rather than empirical investigation because he thinks that it is not an empirical project to determine what are the particular ways that beliefs ought to be manipulated. Whether or not humans reason in the way that they ought to could, insofar as the traditional epistemologist is concerned, be an empirical question.

The descriptive epistemologist, because she denies the normative character of epistemology, is committed to denying the normative character of the theory of how we manipulate beliefs. The descriptive epistemologist might suggest that, instead of assessing human rationality, we can study how we in fact reason. This empirical study will not tell us whether humans reason in the right way, but it will give us an account of how we reason. The descriptive epistemologist would suggest that we should not want anything more than this.

The naturalized epistemologist agrees with the traditional epistemologist that rationality is a normative notion, but thinks that empirical considerations could be relevant to determining what counts as rational. The naturalized epistemologist thinks this because he thinks that to be rational is to manipulate our beliefs in the right way through reasoning and because he thinks that empirical considerations are relevant to how our beliefs ought to be manipulated. It does not bode well for naturalized epistemology if whether or not humans are rational is a conceptual matter. If human rationality is a conceptual issue, then empirical knowledge is not relevant to a major portion of the way we acquire knowledge. Still, naturalized epistemology would not be refuted; empirical evidence might be relevant to ways we acquire knowledge besides reasoning.

Throughout the chapters that follow, I will be engaging in epistemology, more often than not, through an examination of reasoning. In Chapter Seven, I will return specifically to these issues and consider what implications the conclusions of earlier chapters have for epistemology in general. In particular, I will argue these considerations count in favor of naturalized epistemology.

### III. Coming Attractions

Before one can give and defend an answer to the question 'Are humans rational?', she must understand what kind of question this is. In the remainder of this chapter, I sketch various options. Along the way, I offer a guide to the arguments I will be discussing in chapters to come.

The main reason for believing the irrationality thesis is that the results of experiments about reasoning (like the conjunction experiment)—what I call *the reasoning experiments*—show that the principles that characterize human reasoning competence diverge from the normative principles of reasoning. Evidence that humans fail to reason in accordance with the normative principles of reasoning seems to be evidence for the irrationality thesis. For example, experimenters ask subjects whether they think Linda is more likely to be a bank teller or whether she is more likely to be a bank teller and a feminist. If they find that subjects think that the second possibility is more likely than the first (as they in fact do<sup>14</sup>), then they cite this as evidence for the irrationality thesis. This counts as evidence that humans are irrational because it is irrational to believe that a statement **A** is more likely than another statement **B**, which is true whenever **A** is and sometimes when **A** is not. The conjunction experiment will be discussed in more detail in Chapter Three along with another experiment, called the selection task, which investigates our ability to apply rules of deductive logic.

That the reasoning experiments establish the truth of the irrationality thesis is based on the following straightforward way of understanding the nature of the question 'Are humans rational?' In order to figure out whether humans are rational, we start with a criterion for being rational, and we determine whether humans fit this criterion. Without going into any of the details, whether a creature is rational involves the sorts of reasoning

---

<sup>14</sup> Tversky and Kahneman, 'Extensional Versus Intuitive Reasoning'.

processes it uses. To determine, then, whether humans are rational, we must study how humans reason. If humans reason in the right way, then we are rational; if we do not, then we are irrational. To determine how humans reason, we need to look at the world, in particular, at human psychology. Human psychology is an empirical discipline; it involves looking at the behavior and the brains of humans. To study human psychology, we observe human behavior, perhaps by performing experiments like the reasoning experiments, and we look at how humans are constructed, perhaps by cracking open the skull to have a look at the human brain. The details aside, the question of human rationality requires determining how humans reason and this requires doing some psychology, an empirical discipline. If this line of thought is right, then the question of human rationality is empirical and this could be the end of this book, a book so short and trivial that you would be reading it without good reason.

Matters are far from this simple. In reaction to these experiments and the pronouncements that humans are irrational which are based upon them, philosophers and others have defended the claim that humans are rational with a diverse set of interesting and plausible arguments that draw from epistemology, philosophy of science, philosophy of mind, philosophy of language, linguistics, computational theory, and evolutionary theory. In Chapters Three through Seven of this book, I will develop these arguments with an eye towards understanding and evaluating the claim that humans are rational. These arguments try to show that even in light of the reasoning experiments, humans are rational. Most of the arguments for this conclusion are conceptual (the argument about interpretation discussed in Chapter Three and the evolutionary arguments discussed in Chapter Six are the exceptions) and most of them accept the standard picture of rationality (two of the arguments discussed in Chapter Seven are the exceptions). My discussion will also explore the implications of this inquiry to the project of epistemology. Before I map out the structure of the rest of the book, I want to say something about what it means for a question to be empirical or conceptual.

### A. Empirical and Conceptual Matters

An empirical question asks about the particular details of the world we live in. Typically, to answer a question of this sort, one must have a look at the world. To determine how much an electron weighs or how tall my sister is, one must examine the way things are. This is done by performing experiments, taking measurements, and so on. Conceptual questions do not require this sort of examination of the actual world. A conceptual question is answered by reflection on the relevant concepts. To determine whether all bachelors are unmarried, one does not systematically survey all bachelors to see whether or not each of them is married; instead, one can figure out that all bachelors are unmarried by analyzing the concepts involved. A bachelor *just is* an unmarried man; therefore, by definition, all bachelors are unmarried. Part of my project in this book is to determine whether the question 'Are humans rational?' is an empirical question, and, if so, what sort of empirical facts are relevant to it.

Some might argue that this part of the project is based on a mistake because the distinction between empirical and conceptual statements is less clear than it may seem. In his classic article, 'Two Dogmas of Empiricism', Quine discredits some of the traditional assumptions relating sensory experience to the status and truth-values of our beliefs.<sup>15</sup> According to Quine, even such seemingly straightforward empirical statements like 'There are brick houses on Elm Street.' can be insulated from sensory data that would normally be taken to verify or falsify them; such a sentence 'can be held true come what may . . . [for example] by pleading hallucination or by amending . . . logical laws.'<sup>16</sup> If I were taken on a thorough tour of Elm Street and shown that each house I encounter is

---

<sup>15</sup> Willard V. O. Quine, 'Two Dogmas of Empiricism', in *From a Logical Point of View* (Cambridge: Harvard University Press, 1961), 20-46.

<sup>16</sup> *Ibid.*, 43.

made of wood, I could still, if I wanted to, hold on to the statement 'There are brick houses on Elm Street.' by insisting that I must have missed a house or that I must have mistaken as wooden at least one house that is in fact made of brick. An ingenious person can take any statement that *prima facie* seems to require looking at the details of the world and show that it actually does not require such evidence. Similarly, even those statements that seem completely immune from empirical refutation, might be revised in the face of some empirical evidence. If this is right, then the distinction between empirical and conceptual statements may be in trouble. Above I characterized a statement as empirical if its truth depends on the way the world is. If Quine is right that the truth of 'There are brick houses on Elm Street.' may not depend on the way the world is and that statements like '**A and B** implies **A**.' may be rejected if the world turns out in a particular way, then the question of whether human reasoning is conceptual or empirical may be uninteresting.

A response to this line of thought would be to try to salvage the conceptual-empirical distinction by appeal to Quine's own idea that beliefs can be thought of as located in conceptual space at 'varying distances from the sensory periphery'.<sup>17</sup> The idea is to see each person as having a 'web of belief',<sup>18</sup> an interlocking set of beliefs in which beliefs that are revisable in the face of experience are close to the border of the web and beliefs that are highly unlikely to be revised in the face of experience are in the center of the web. The idea would be to recast the question about human rationality as follows: is the belief that humans are rational likely to be revised in the face of evidence, in particular, evidence concerning how humans in fact reason? Someone who thinks that the reasoning experiments can do nothing to undermine our belief in human rationality would see the

---

<sup>17</sup> Ibid.

<sup>18</sup> This idea is developed in Willard V. O. Quine and J. L. Ullian, *The Web of Belief* (New York: Random House, 1970).

belief that humans are rational as being in the center of the web, while someone who thinks the reasoning experiments might have 'bleak implications' for human rationality would see the belief that humans are rational as close to the edge of the web.

This attempt to recast the debate about the nature of the question whether humans are rational might seem vulnerable to a deeper Quinean worry. Quine does not simply substitute "near the edge of the web" for "empirical" and "in the center of the web" for "conceptual"; the point of his invocation of the web is that beliefs we take to be empirical can be held constant come what may and beliefs we take to be conceptual can be revised in the face of certain kinds of evidence. For example, in pre-Copernican times, the belief 'The earth is the center of the universe.' was in the center of most people's webs. Over time, this belief, which seemed irrevisable, was rejected in the face of astrological and other scientific evidence.<sup>19</sup> If this is right, then it might seem that there is nothing special about a belief that is currently at the center of most people's webs of belief beyond the fact that such a belief is *at present* well-insulated from empirical refutation. This gives rise to the deeper worry that the issue of whether the belief 'Humans are rational.' is at the periphery or in the center of the web is roughly equivalent to a Gallup poll of how strongly people hold the belief that humans are rational and thus would at best be a matter more for sociology rather than philosophy or psychology.

I want to suggest that this deeper worry need not be taken very seriously, even if one is persuaded to reject the standard distinction between empirical and conceptual truths. While Quine does argue that there is no such thing as an intrinsically irrevisable belief or an intrinsically 'revisable-in-the-face-of-sensory-evidence' belief, he does not argue for and is not committed to the view that a belief's status (as irrevisable or revisable by sensory experience) is arbitrary or unimportant. In fact, since Quine is a holist, namely,

---

<sup>19</sup> Thomas Kuhn, *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought* (Cambridge: Harvard University Press, 1957).

he thinks that our beliefs about the world 'face the tribunal of sense experience not individually but as a corporate whole,'<sup>20</sup> he sees all of our various beliefs as extensively interconnected and their location in the web of belief as dependent on the sorts of interconnections that hold among them. If this is right, then where the belief 'Humans are rational.' fits in the web is constrained by its connections to other beliefs and their position in the web. Further, given that humans share roughly the same sensory apparatus and roughly the same principles of belief revision, it follows that the placement of a particular belief in the web is not at all arbitrary or uninteresting. When I inquire whether 'Humans are rational.' is a conceptual or empirical statement, according to the web of belief picture, I am inquiring about how beliefs about human rationality relate to other beliefs, in particular beliefs about human psychology and neurophysiology and beliefs about epistemology, philosophy of language and the like. If the arguments against the traditional distinction between conceptual and empirical facts are strong, then the question about the status of human rationality might be less clear, but it remains of interest insofar as we are interested in the relations among the various disciplines and among different kinds of knowledge. I am not here embracing the Quinean criticism of the conceptual-empirical distinction or his 'web of belief' picture. My point is that even if this criticism is right, assessing the nature of the question of human rationality is still a worthwhile project.<sup>21</sup>

Even with a clear picture of the distinction between what is conceptual and what is empirical, it is not at all clear what to say about the question 'Are humans rational?' There are two kinds of arguments to the effect that the reasoning experiments are not relevant to the issue between the rationality thesis and the irrationality thesis: arguments

---

<sup>20</sup> Quine, 'Two Dogmas', 41.

<sup>21</sup> A more extensive discussion of some of these issues in Quine can be found in Christopher Hookway, *Quine: Language, Experience and Reality* (Stanford: Stanford University Press, 1988), Chapter Two.

that accept the standard picture of rationality and arguments that reject this picture. I discuss diverse versions of each of these arguments in turn.

#### B. Arguments for the Rationality Thesis within the Standard Picture

The first two arguments for the rationality thesis that accept that the standard picture of rationality involve issues about how to interpret human behavior and human cognitive mechanisms in general and the reasoning experiments and human reasoning competence in particular. In Chapter Three, I discuss a general interpretative strategy of importance to friends of the rationality thesis. The strategy is to interpret every instance of a person failing to reason in accordance with the normative principles of reasoning as a performance error. This interpretive strategy, if justified, would have some significant ramifications for my inquiry. The first ramification would be that, so interpreted, the results of the reasoning experiments would fail to suggest that humans are irrational. Second, without the evidence of the reasoning experiments, the irrationality thesis would lose its primary source of support and, as a result, the rationality thesis would be in good shape. The third ramification would be that empirical considerations would not be relevant to the question of human rationality because this interpretive strategy would discount any evidence in favor of human irrationality. Note that this argument is empirical; it implies that, because of the many performance errors we make, we can never have access to human reasoning competence. In Chapter Three, I explain why this strategy has these ramifications and why this strategy must be justified before it can be used.

In Chapter Four, I consider an argument based on the principle of charity. The principle of charity is a guide for translating utterances of other speakers; the idea is that

translation should be charitable to the person being translated.<sup>22</sup> This principle can be extended to the interpretation of people's reasoning competence and more generally to the interpretation of people's cognitive mechanisms.<sup>23</sup> One general argument for the principle of charity is that you must assume that the person you are trying to understand is at least somewhat rational, because you will not be able to make sense of a truly non-rational person. The principle of charity argument for the rationality thesis is that cognitive scientists must be mistaken in interpreting subjects in the reasoning experiments as being irrational because doing so would violate the principle of charity. Interpreters of the reasoning experiments (including the experimenters who perform them) attribute non-normative principles of reasoning to subjects and, based on these experiments, to humans in general. The principle of charity argument for the rationality thesis says that such attributions are mistaken: the principle of charity should be applied to subjects in the reasoning experiments. Properly done, this would show that the reasoning experiments are fully consistent with the rationality thesis and do not provide any support for the irrationality thesis. I distinguish between a strong and weak version of the principle of charity applied to people's reasoning competence—the weak version of the principle of charity is defeasible (it says that people should be interpreted as rational *unless* there is strong evidence to suggest otherwise) while the strong version of the principle of charity is not (it says that people should *always* be interpreted as rational). I argue that the weak principle of charity is justified, but it does not provide an argument

---

<sup>22</sup> Willard V. O. Quine, *Word and Object* (Cambridge: MIT Press, 1960); and 'Ontological Relativity', in *Ontological Relativity and Other Essays* (NY: Columbia University Press, 1969).

<sup>23</sup> 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), 165-91; and Cohen, 'Can Human Irrationality Be Experimentally Demonstrated?'

for the rationality thesis. The strong principle of charity, if justified, would provide an argument for the rationality thesis, but I argue that it is not justified.

The remaining arguments for the rationality thesis that accept the standard picture of rationality are not specifically about the interpretation of the reasoning experiments. In Chapter Five, I turn to the *reflective equilibrium* argument for the rationality thesis. The general idea behind this argument is that since both our norms of reasoning and our actual reasoning behavior are based on our intuitions about what counts as good reasoning, reasoning behavior and the norms cannot diverge; insofar as the reasoning experiments seem to prove that reasoning behavior diverges from the norms, they do so only in virtue of detecting the performance errors we make.<sup>24</sup> The most interesting version of this argument involves the theory of reflective equilibrium, an epistemological theory that says a set of principles is justified when it is modified to fit with first-order intuitions about its domain of application.<sup>25</sup> The idea is that both the normative principles of reasoning and the description of reasoning competence come from a process of reflective equilibrium with our intuitions about what counts as good reasoning as input. As such, the two sets of principles cannot diverge. This is a conceptual argument for the rationality thesis; if it succeeds, the question of human rationality is not empirical.

---

<sup>24</sup> Cohen, 'Can Human Irrationality Be Experimentally Demonstrated?'; Macnamara, *Border Dispute*; and Sober, 'Psychologism'.

<sup>25</sup> The concept of reflective equilibrium comes from Nelson Goodman, *Fact, Fiction and Forecast*, fourth edition (Cambridge: Harvard University Press, 1983), 63-64; it was taken up in John Rawls, *A Theory of Justice* (Cambridge: Harvard University Press, 1971), and baptized 'reflective equilibrium' in John Rawls, 'The Independence of Moral Theory', *Proceedings and Addresses of the American Philosophical Association* 48 (1974-75), 5-22.

This argument is usually criticized by saying that the normative principles of reasoning are not the result of a process of reflective equilibrium.<sup>26</sup> I defend the reflective equilibrium account of where the norms come from; I suggest that critics of this account underestimate its resources. The reflective equilibrium argument for the rationality thesis fails, I think, for another reason. Even if the empirical process of giving a complete description of human reasoning competence is a process of reflective equilibrium, the inputs to a reflective equilibrium process to determine our reasoning competence would differ from the inputs to such a process to determine the norms of reasoning. Certain kinds of biological, computational, and neuroscientific evidence are relevant to developing an account of our reasoning competence, but not to determining what the norms of reasoning are. Further, even if the input to the two reflective equilibrium processes were the same—perhaps because epistemology should be naturalized and thus scientific evidence would be relevant to the normative principles of reasoning—different parts of the input would be weighted in different ways as part of the two reflective equilibrium processes. Reflective equilibrium may be the right process for determining what the norms of reasoning are, but the reflective equilibrium argument for the rationality thesis still fails.

The final argument for the rationality thesis that accepts the standard picture of rationality is not, like most of the arguments that precede it, a conceptual argument. The argument says that evolutionary theory, but not the reasoning experiments, is relevant to settling the issue between the rationality and the irrationality theses. The idea of this argument, which I consider in Chapter Six, is that evolution, through natural selection, produces organisms with mechanisms that select true beliefs (what I call *truth-tropic mechanisms*) and that organisms with such mechanisms would be rational ones. Humans, being the result of natural selection, have truth-tropic cognitive mechanisms and such

---

<sup>26</sup> For example, Stich, *Fragmentation*, Chapter Four.

mechanisms give rise to a reasoning competence that instantiates truth-tropic principles of reasoning. This argument provides a *prima facie* reason for not interpreting the results of the reasoning experiments as showing that humans are irrational. The evolutionary argument is supposed to show that the reasoning experiments *cannot* establish the truth of the irrationality thesis because the evolutionary history of our reasoning competence provides a strong reason for interpreting it as rational.<sup>27</sup> This argument has an important empirical premise—the truth of the theory of natural selection—but it has an affinity with conceptual arguments for the rationality thesis because the evolutionary argument says that, given a rather well-established empirical premise, the nature of the question of human rationality does not require any further examination of the world. From the point of view of psychologists who perform the reasoning experiments, for example, the result of the evolutionary argument for the rationality thesis is basically the same as either the reflective equilibrium argument for the rationality thesis or the principle of charity argument for the rationality thesis—namely, the reasoning experiments are irrelevant to the rationality thesis.

---

<sup>27</sup> Daniel Dennett, 'Making Sense of Ourselves', in *Intentional Stance*, 83-101; Jerry Fodor, 'Three Cheers for Propositional Attitudes', in *Representations* (Cambridge: MIT Press, 1981), 100-23; Alvin Goldman, *Epistemology and Cognition* (Cambridge: MIT Press, 1986); William Lycan, 'Epistemic Value', in *Judgment and Justification* (Cambridge: Cambridge University Press, 1988), 128-56; Ruth Millikan, 'Naturalist Reflections on Knowledge', *Pacific Philosophical Quarterly* 65 (1984), 315-34; David Papineau, *Reality and Representation* (Oxford: Basil Blackwell, 1987); Karl Popper, 'Evolutionary Epistemology', in *Evolutionary Theory: Paths into the Future*, Jeffrey Pollard, ed. (London: Wiley and Sons, 1984), 239-56; W. V. O. Quine, 'Natural Kinds', in *Ontological Relativity and Other Essays*, 114-38, reprinted in *Naturalizing Epistemology*, Kornblith, ed., 57-76; and Elliott Sober, 'Evolution of Rationality', *Synthese* 46 (1981), 95-120.

The evolutionary argument for the rationality thesis has typically been criticized for adopting a simplistic picture of evolution.<sup>28</sup> Steven Jay Gould and Richard Lewontin use the adjective "Panglossian" to describe the knee-jerk practice of appealing to natural selection as the evolutionary force that explains every trait; the term comes from Voltaire's Dr. Pangloss who said, for example, that the reason humans have noses is to hold up our eyeglasses.<sup>29</sup> Gould and Lewontin correctly point out that there are other forces of evolution besides natural selection. A critic of the evolutionary argument for the rationality thesis could point out that these forces might be behind the evolution of our reasoning competence. If this is the case, then there is no guarantee that our reasoning competence will be truth-tropic; the evolutionary argument for the rationality thesis would thus fail. I defend the evolutionary argument against this criticism. Natural selection is the only evolutionary force that can explain functionally complex structures.<sup>30</sup> Since our reasoning competence is functionally complex, if it evolved, then it did so through natural selection.

The evolutionary argument for the rationality thesis fails, I think, for another reason. Even though our reasoning competence must be the result of natural selection, the operation of natural selection does not guarantee that all of our principles of reasoning will be truth-tropic. The criterion for natural selection is reproductive success and sometimes being truth-tropic and leading to reproductive success come apart. Because

---

<sup>28</sup> For example, Stich, *Fragmentation*, Chapter Three.

<sup>29</sup> 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), 281-8, reprinted in *Conceptual Issues in Evolutionary Biology*, second edition, Elliott Sober, ed. (Cambridge: MIT Press, 1994), 73-90.

<sup>30</sup> George C. Williams, *Adaptation and Natural Selection* (Princeton: Princeton University Press, 1966).

natural selection cannot guarantee truth-tropicity, natural selection cannot guarantee rationality; the evolutionary argument for the rationality thesis thus fails.

There are two 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 evolution involves selection for reproductive success and the claim that having principles of reasoning 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 principles of reasoning that lead to reproductive success and, hence, they will be rational. This argument only works by rejecting the standard picture of rationality; maximizing reproductive success is rational in a very different sense than reasoning in accordance with rules of logic.

The second modification to the evolutionary argument for the rationality thesis draws from an approach to theory of knowledge known as evolutionary epistemology.<sup>31</sup> The two versions of the evolutionary argument considered thus far involve innate mental mechanisms. But perhaps the principles that guide reasoning are *not* innate.<sup>32</sup> If this is the case, then biological evolution and natural selection cannot be the driving forces behind the development of our reasoning competence. This is where evolutionary

---

<sup>31</sup> See, for example, Donald Campbell, 'Evolutionary Epistemology', in *The Philosophy of Karl Popper*, volume 1, Paul Schilpp, ed. (LaSalle, IL: Open Court, 1974), 413-63; Michael Bradie, 'Assessing Evolutionary Epistemology', *Biology and Philosophy* 1 (1986), 401-59; Bradie, 'Epistemology from an Evolutionary Point of View', in *Conceptual Issues in Evolutionary Biology*, second edition Sober, ed., 453-76; and Edward Stein, 'Evolutionary Epistemology', in *A Companion to Epistemology*, Jonathan Dancy and Ernest Sosa, eds. (Oxford: Basil Blackwell, 1992), 122-5.

<sup>32</sup> See, for example, Stich, *Fragmentation*, 71-4, for arguments on this point. I criticize such arguments in Chapter Two below.

epistemology is supposed to be relevant. According to this view, the development of human knowledge is governed by a trial-and-error process analogous to biological natural selection. If the principles 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 principles that produce true beliefs; having such truth-tropic principles makes an organism rational; because, according to evolutionary epistemology, humans acquire their principles of reasoning through epistemic natural selection, humans are therefore rational. The argument for human rationality based on conceptual evolution fails because natural selection will not select only truth-tropic principles of reasoning. This is similar to the reasons the argument based on biological evolution failed, but the conceptual evolution argument fails independent of the details of biology. This shows that even the general structure of evolutionary theory does not guarantee rationality.

### C. Arguments for the Rationality Thesis That Reject the Standard Picture

For the irrationality thesis to be true, the normative principles of reasoning and human reasoning competence must diverge. The preceding arguments for the rationality thesis accepted the standard picture of rationality and tried to show that human reasoning competence does not diverge from the normative principles of reasoning. Friends of the rationality thesis might proceed in another manner: rather than argue that the reasoning experiments are not about reasoning competence, they can attack the irrationality thesis in the other direction, by arguing that the standard picture of rationality is mistaken with respect to the normative principles of reasoning. Assuming, for example, that the rationality thesis is wrong and that the conjunction principle is not a normative principle of reasoning, subjects in the Linda experiment may well be reasoning in accordance with the norms. If we are mistaken in our account of what the normative principles of

reasoning are, then humans might turn out to reason in accordance with the norms of reasoning and, hence, be rational even in the face of the reasoning experiments. In Chapter Seven, I consider three attempts to undermine the irrationality thesis in roughly this way.

For humans to inquire whether humans are rational is a quite different thing than for us to inquire what the mass of an electron is or whether humans are mammals. Whether humans are rational may be more like the question of whether humans have a good aesthetic sensibility and the question of whether humans have a good sense of humor. Suppose, for example, experimenters showed subjects works of art (some beautiful and some not beautiful), asked them to say whether or not each work of art is beautiful or not, and, on the basis of whether the subjects correctly distinguished the beautiful works of art from the non-beautiful ones, thereby determined whether or not humans have a good aesthetic sensibility. Or suppose that experimenters told subjects some jokes (some funny and some not), asked subjects to identify whether each joke was funny or not, and, on the basis of whether the subjects correctly distinguished the funny from the unfunny jokes, thereby determined whether humans have a good sense of humor. What do you think of these proposed experiments? Would such experiments actually determine whether humans have a good aesthetic sensibility or a good sense of humor? It seems quite odd to say that they would. In fact, it is tempting to say that whether or not humans have a good aesthetic sensibility or a good sense of humor is not the sort of question about which an empirical inquiry can be undertaken.

A few different intuitions are at work here. First, these experiments assume that there are standards of funniness or aesthetic valuation that apply to all humans, but it is not obvious that there are any such general standards. Relativism, the view that the standards in some realm are relative to an individual—or a small group of people—is true for some realms, for example, ice cream flavors. When I say that Fudge Swirl is the best kind of ice cream, I mean that it is the best ice cream from my point of view, *not* that it is true for

all humans. Roughly the same may be true with respect to sense of humor and aesthetic evaluation. When I say that the joke about what you get when you cross a Mafioso and a literary theorist (someone who makes you an offer you can't understand) is funny, I mean that *I* think it is funny, but I do not mean that everyone else ought to find it funny. If I tell the joke, this indicates that I suspect that some other people might think it is funny, but I allow that there are others who have a sense of humor that differs from mine. On some views, the same is true for aesthetic valuations. When I say that a particular work of art is beautiful, I am just expressing my own opinion. I leave open the possibility that others will disagree with me and that their opinion could be as reasonable as mine. Relativism is true for ice cream flavors and at least plausible with respect to funniness and aesthetics. If relativism is true for funniness, then the experiment I describe above would rest on a mistake: the experiment mistakenly assumes that it is an empirical question whether humans have a good sense of humor. In fact, the question may be conceptual. Each of us may have our own sense of humor and, associated with it, our own standard of funniness.

Another reason why the proposed experiments about aesthetic sensibility and sense of humor seem odd is that such experiments assume there is some standard of aesthetic valuation (or funniness) that exists independent of the human aesthetic (or 'amusement') faculty. It seems plausible, however, that even if there are standards that apply to all humans, there are no standards that exist independent of humans because the only standards that exist are indexed to us. Our human aesthetic faculty makes it the case that the features that make things seem beautiful to us—say, certain color combinations, the right mix of unity and variety, and so on—are in fact the features that make things beautiful. Things that seem beautiful to us are unlikely to seem beautiful to creatures with sensory faculties dramatically different to ours. If this is right, then no experiment could possibly show that humans have a poor aesthetic sensibility because the standards of aesthetic evaluation come from our own aesthetic capacities. Further, no empirical

investigation into our particular aesthetic faculties is required to determine that humans have a good aesthetic sensibility because the standards of aesthetic valuation come from our aesthetic capacities. This shows that humans have a good aesthetic sensibility and it does so independent of any empirical facts about our particular aesthetic capacities because the normative standards come from whatever aesthetic capacities we have.

Finally, even if there is a standard of aesthetic valuation (or funniness) that is independent of human capacities and that applies to all humans, the proposed experiments might seem odd because they are based on the assumption that experimenters have access to which works of art are and are not beautiful and that they can use this knowledge to assess the aesthetic abilities of the subjects. But what gives the experimenters this special access? They could decide for themselves that artworks **A**, **B** and **C** are beautiful and artworks **D**, **E**, and **F** are not or they could consult art experts to make this sort of determination. But on what basis do the experimenters' or the experts' assessments count as the *right* judgments of what is beautiful while the subjects' assessments do not? Further, if we grant this special status to the experimenters or the art experts, we thereby *assume* that humans (at least some of us) have good aesthetic judgment. The problem is that even if there are standards of aesthetic valuation that are independent of us, there is no guarantee that we have access to such standards. Without such a guarantee, experimenters are not able to determine whether the principles that guide our aesthetic judgment are the right principles for aesthetic judgment. Empirical considerations will not help matters—even if we can determine what aesthetic principles we follow, we are in no position to assess them against the standards of what is beautiful. The experiment to determine whether humans have a good aesthetic sensibility seems odd because such an experiment depends on our having access to the very standards that the experiment is supposed to tell us whether or not we can access.

There are thus three worries that arise with respect to the proposed experiments to determine whether humans have a good aesthetic sensibility and whether humans have a good sense of humor:

- (1) There might be no objective or general standards of aesthetic valuation (or funniness)—that is, beauty (and funniness) might be relative to the tastes of individual humans and, thus, there might be no general standards of beauty (or funniness).
- (2) There might be objective standards of aesthetic valuation (or funniness), but these standards might be indexed to human faculties—that is, there might be no standards of beauty (or funniness) that are independent of humans.
- (3) Although there might be general, human-independent standards of aesthetic valuation (or funniness), we might not have access to them.

To return to reasoning, many philosophers and some other people are suspicious of those who claim that the reasoning experiments are relevant to the question of whether humans are rational, and, more generally, whether any empirical considerations are relevant to human rationality. Some of this suspicion comes from the same sort of intuitions that counted against the imagined experiments concerning aesthetic judgment. In particular, whether or not we are rational would not be straightforwardly empirical if any of the following situations held true:

- (1) Relativism is true about reasoning. There are no normative principles of reasoning that apply to all humans. Whether a principle counts as good for reasoning is indexed to each person.
- (2) What counts as good reasoning is indexed to human reasoning ability in general; there are no normative principles of reasoning independent of our reasoning abilities against which to compare these abilities.

- (3) There are normative principles of reasoning independent of our reasoning abilities but we have no way of getting outside of our own reasoning faculties to compare them to the normative principles of reasoning.

Each of these possibilities will be discussed at length in the course of this book. For now, I want to say something brief about each.

First, consider the possibility that there are no general standards of what counts as good reasoning that apply to humans in general. While relativism is plausible in certain realms (taste, funniness, and perhaps aesthetics), *prima facie*, relativism seems implausible with respect to reasoning. It seems crazy to say that reasoning in accordance with principles based on rules of logic is a good thing for some people but not for others. Violating principles based on rules of logic seems wrong for everyone, not just for some folks. (1) is false; therefore, it fails to show that human rationality is not an empirical issue. Relativism is not, however, this straightforwardly implausible with respect to reasoning.

One way to make relativism about reasoning more plausible is to point to the finite computational resources that each human has. We each have a limited amount of memory, a limited life span, a limited-sized brain and a limited amount of resources to devote to the project of reasoning.<sup>33</sup> Given these limitations, we need to develop *efficient* reasoning strategies. Different people, because of their different intellectual resources, their different interests, their different needs, and their different environments will develop and follow different rules of reasoning. Looked at this way, relativism about the principles of reasoning is not so implausible.

The truth of this sort of relativism does not entail that the question 'Are humans rational?' is a conceptual question. How many calories a person should consume in a day

---

<sup>33</sup> Christopher Cherniak, *Minimal Rationality* (Cambridge: MIT Press, 1986), discusses the implications of the human finitary predicament for issues related to human rationality.

is indexed to her height, her age, her rate of metabolism, and the like. This shows that different people should consume different amounts of calories, but it does not prove that how many calories a person should eat is a conceptual matter. For relativism about human rationality to entail that human rationality is a conceptual issue, one must show either that there are no normative principles of reasoning at all (the existence of the human finitary predicament does not establish this—relativism does not entail nihilism) or one must show that the way a person should reason is necessarily the same as how that person in fact reasons. The truth of relativism would entail that the question 'Are humans rational?' is a conceptual question only if there was some further argument to show that all of the divergences that a person makes from the principles that are the normative principles of reasoning for that person must be performance errors. This argument has affinities to the argument (sketched above and discussed in Chapter Three) that accepts the standard picture of rationality and says that all divergences from the normative principles must be performance errors. Just as the Chapter Three argument fails, so too, I argue in Chapter Seven, does the relativistic argument for the rationality thesis.

Second, consider the possibility that the normative principles of reasoning are indexed to human reasoning ability in general. If this is the case, it is hard to see how we could fail to reason in accordance with the norms because the norms are based on us. If the normative principles of reasoning just come from our reasoning capacity, then humans must be rational. If this is right, then human rationality is a conceptual question, and the answer to the question is that we *are* rational. This argument might seem similar to the reflective equilibrium argument that I discuss in Chapter Five, but there is an important difference. The reflective equilibrium argument accepts the standard picture of rationality and attempts to show that our normative principles of reasoning must be the same as the principles embodied in our reasoning competence. The present argument attempts to reach the same conclusion but by rejecting the standard picture of rationality. In Chapter Seven, I argue that this attempt fails.

Finally, consider the possibility that we have no way of getting outside our reasoning faculties. In the case of the proposed experiment about human aesthetic sensibilities, the problem was how the experimenters could determine for themselves which artworks were beautiful and which were not. The parallel problem with respect to reasoning is that experimenters who wish to determine whether humans are rational need to start with an account of which principles of reasoning are rational and which ones are not. Unless they already assume that they themselves are rational, they will be unable to do so.

There is a more specific version of this possibility that applies to reasoning. In order to inquire whether we are rational, we need to use our reasoning faculties, the very faculties that we are trying to assess. The problem is this: suppose we are irrational, that is, that our reasoning faculties cannot be relied upon to work properly. If this is the case, when we inquire into the matter of whether or not we are rational, we cannot be relied on to get the right answer; even if we conclude that we are rational, we might well be wrong, because at least some of our methods for reaching this conclusion are, by stipulation, irrational. Now suppose, as is the case, that we do not know whether we are rational or not and, further, that we want to determine whether or not we are. If we are truly in doubt as to our reasoning abilities, then we cannot trust the results of our inquiry because if we are irrational, then our methods of inquiry may be unreliable. This suggests that the irrationality thesis cannot be established by empirical methods. It does not show, however, that the issue of human rationality is a conceptual one. A question can be empirically unanswerable yet not conceptual if an answer to the question is empirically inaccessible. An example of this is the problem of other minds, namely, how can I know there are minds in the world besides my own when the only evidence I have for other minds is based on the behavior of other bodies? For all I can observe, everybody else in the world might be a robot. According to this analysis, it is possible that there are other minds, but the question may not be empirically answerable because there might not be any empirical evidence that I could have that would bear on the question of whether there

are other minds. The question of whether humans are rational might also be an empirical question with an epistemologically inaccessible answer. This is consistent with the suggestion that we cannot get outside of our reasoning abilities. The first possibility thus does not establish that the issue of human rationality is a conceptual question, although it does undermine the idea that human rationality is a straightforwardly empirical matter. I will criticize this argument in Chapter Seven.

#### D. Assessing the Standard Picture

These three possibilities—there might be no normative principles of reasoning that apply to all humans; there might be such principles, but they might not be independent of our reasoning abilities; and there might be such independent principles, but we might not be able to compare our reasoning abilities against some extra-human normative principles of reasoning—attempt to defend the rationality thesis by undermining the standard picture of rationality. In Chapter Seven, I argue that they fail to establish the rationality thesis. This leaves open the possibility that, even though they do not successfully establish the rationality thesis, these arguments successfully undermine the standard picture of rationality, and that some alternative pictures of rationality—for example, what I call the *pragmatic picture of rationality* and the *relativistic picture of rationality*—might be the proper account of what rationality is.

In the rest of Chapter Seven, I turn my attention to the standard picture of rationality. Up to this point, I will have assumed that the standard picture of rationality is true but I will not have said much in its defense. In Chapter Seven, I will discuss the virtues of the standard picture of rationality and consider two arguments against it. The first argument has to do with the human finitary predicament. The worry is that some of the principles of reasoning that are deemed norms by the standard picture of reasoning are not feasible for humans to reason in accordance with. If so, then the standard picture of rationality must be mistaken because an unfeasible principle cannot be a norm. The second

argument accuses friends of the standard picture of rationality of what Stephen Stich calls *epistemic chauvinism*.<sup>34</sup> The idea is that the only reason we have for embracing the standard picture's account of what the normative principles of reasoning are is that these principles are favored in our culture, our language and our way of thinking. This is not, however, a good reason for accepting an account. I develop responses to both of these arguments, but the standard picture is, I think, still in trouble. In the remaining sections of Chapter Seven, I develop an alternative to the standard picture that I call the *naturalized picture of rationality*. According to this picture, various empirical facts about humans and our environment must be taken into consideration in determining what the normative principles of reasoning are. I argue that this account has the virtues of the standard picture of rationality while avoiding the problems that face the standard picture.

#### E. The Punch Line

In Chapter Eight, I conclude my discussion by returning directly to the central questions that began this inquiry: 'Are humans rational?', 'Is this a conceptual or empirical question?', and 'If this is an empirical question, what kind of evidence will it answer?' I here explain how the preceding discussion establishes that the question whether humans are rational is primarily an empirical question, the sort of empirical question that the reasoning experiments help answer. Further, I suggest that the reasoning experiments give us some reason to think that humans are irrational, although there is more empirical work to be done, not just in psychology proper, but in evolutionary theory, computational theory, and neuroscience. My conclusion here is based on the naturalized picture of rationality developed in Chapter Seven. I conclude by discussing the broader morals of this inquiry, in particular to project of naturalizing epistemology.

---

<sup>34</sup> Stich, *Fragmentation*, 94, and *passim*.