The Oxford Handbook of Rationality

Mele, Alfred R. (Editor), Professor of Philosophy, Florida State University
Rawling, Piers (Editor), Professor of Philosophy, Florida State University

2004

Chapter 1
INTRODUCTION

Aspects of Rationality

Alfred R. Mele
Piers. Rawling

This volume consists of two main parts. The first examines the nature of rationality broadly understood. The second explores rationality's role in and relation to other domains of inquiry: psychology, gender, personhood, language, science, economics, law, and evolution. Our aim in this introductory essay is to sketch the theoretical terrain on which this volume is situated and to introduce the subsequent chapters.

1. The Nature of Rationality

The domain of rationality is customarily divided into the theoretical (see Robert Audi's chap. 2) and the practical. Whereas theoretical or epistemic rationality is concerned with what it is rational to believe, and sometimes with rational degrees of belief, practical rationality is concerned with what it is rational to do, or intend or desire to do. In this section, we raise some of the main issues relevant to philosophical discussion of the nature of rationality and then briefly describe the chapters in part 1.
One obvious issue concerns the relation between practical and theoretical rationality. Discussions of the nature of practical rationality and reason concern norms of choice, and it seems that if such norms are not arbitrary, arguments over what those norms are must ultimately be a theoretical matter. To suppose otherwise would seem to generate an infinite regress: if we could choose norms of choice on a rational basis, then this rational basis would itself require norms chosen on a rational basis, and so on. This issue arises, at least implicitly, in David Gauthier's approach, which is discussed by James Dreier in chapter 9. Conversely, practical considerations enter into the theoretical domain. This is examined by Gilbert Harman in chapter 3; and in one of the phenomena Alfred Mele explores in chapter 13—motivationally biased belief—practical considerations sometimes seem to influence beliefs in ways that violate epistemic norms (also see Samuels and Stich, chap. 15).

Harman explicitly discusses reasoning. What is the relation of reasoning to rationality? On certain decision-theoretic approaches (see James Joyce's chap. 8 and James Dreier's chap. 9), for example, rationality requires only that one's preferences meet certain ordering criteria: nothing is said about processes of reasoning about preference. In particular, decision theory does not require explicit calculation of expected utilities. Decision theory is one approach in which rationality is seen as a matter of internal consistency. Minimally, the idea behind internal consistency approaches to rationality is that one might be rational and yet have false beliefs and perverse preferences provided that one is in some sense coherent (see chap. 4 by Brad Hooker and Bart Streumer, chap. 5 by Michael Smith, and chap. 7 by David McNaughton and Piers Rawling for more on internal consistency approaches). Although Hume (see Smith, chap. 5) does not actually use the term "rational," he is the historical figure perhaps most often associated with the idea that perverse preferences can be rational. And Kant is perhaps the figure most often associated with the denial of this, at least for the class of perverse preferences that motivate immoral action. In chapter 6, Onora O'Neill presents a Kantian argument for the claim that it is irrational to be immoral. (Related issues on Humean themes are whether beliefs can, by themselves, rationally require certain motives, and whether beliefs can, by themselves, produce those motives. See chaps. 5 and 7.)

Sometimes the issue of the basis of morality is put in terms of reasons: does one have reason to be moral? If it is supposed irrational to fail to do what you have most reason to do, this question is closely related to that of whether rationality prescribes doing as morality requires. But some authors deny that rationality requires doing what you have most reason to do: one might have an internal-consistency view of rationality but regard reasons as a more "external" or "substantive" matter (see chaps. 4, 5, and 7). Related to this is the vexed question of whether you have a reason to *A* only if you desire to *A,* or could reach such a desire if you were to reason in some appropriate fashion. (See chaps. 4 and 7.)

Among other issues addressed by the authors in part 1 are the relations between rationality and the emotions (Patricia Greenspan, chap. 11), the rationality of being guided by rules (Edward McClennen, chap. 12), the nature and causes of irrationality (Alfred Mele, chap. 13), and paradoxes of rationality (Roy Sorensen, chap. 14).

No contemporary discussion of rationality would be complete without significant material on the use of formal methods in its study. James Joyce examines Bayesianism as a unified theory of epistemic and practical rationality in chapter 8, with a focus on Bayesian epistemology. James Dreier, in chapter 9, shows how the formal apparatus of decision theory is connected to some abstract issues in moral theory. And the use of game theory to model interaction between decision makers is the topic of Cristina Bicchieri's chapter 10.

We turn now to summaries of the chapters in part 1.

In “Theoretical Rationality: Its Sources, Structure, and Scope” (chap. 2), Robert Audi presents an account of the nature and chief varieties of theoretical rationality, conceived mainly as the rationality of cognitions—especially, beliefs. Audi describes the essential sources of theoretically rational cognitions: perception, memory, consciousness, reason, and testimony. He also examines the role of coherence in accounting for rational belief and distinguishes the evidential and conceptual roles of coherence. In the light of his account of sources of belief and knowledge, Audi describes the structure of a rational system of cognitions in persons whose beliefs reflect both direct responsiveness to basic sources of cognition—such as perception—and inferences that build on those sources. He considers conditions for rational change of belief, and he sketches structural and developmental aspects of a person's theoretical rationality. In his concluding sections, Audi discusses the scope of theoretical rationality and the kind of cognitive integration it requires.

In “Practical Aspects of Theoretical Reasoning” (chap. 3), Gilbert Harman distinguishes between two uses of the term “logic”: as referring either to the theory of implication or to the theory of reasoning, which are quite distinct. His interest here is the latter. Reasoning is a process that can modify intentions and beliefs. To a first approximation, theoretical reasoning is concerned with what to believe and practical reasoning is concerned with what to intend to do, although it is possible to have practical reasons to believe something. Practical reasoning differs from theoretical reasoning in allowing arbitrary decisions and a certain sort of wishful thinking. Practical considerations are relevant to whether to engage in theoretical inquiry into a given question, the extent of time and other resources to devote to such inquiry, and whether and when to end such inquiry.

Simplicity and conservatism play a role in theoretical reasoning that can be given a practical justification without allowing wishful thinking into theoretical reasoning, a justification that can also be given a nonpractical interpretation.

Brad Hooker and Bart Streumer, in “Procedural and Substantive Practical Rationality” (chap. 4), distinguish the two thus: according to *proceduralism* an agent is open to rational criticism for lacking a desire only if she fails to have a desire that she can rationally reach from her beliefs and other desires, whereas according to *substantivism* an agent is open to such criticism not only if her desires fail procedurally, but also if they fail substantively—where, for example, an agent who lacks the desire to take curative medicine might be substantively irrational in virtue of this lack, and yet be procedurally rational because she cannot rationally reach this desire from her beliefs and other desires. Hooker and Streumer discuss the proceduralist views of Hume (1739), Brandt (1979 , 1989), and Williams (1981 , 1995a , 1995b), before turning to substantivist arguments. They conclude by noting the advantages of following Scanlon (1998) in being a proceduralist about practical rationality but a substantivist about practical reasons.

In “Humean Rationality” (chap. 5), Michael Smith focuses on the relationship between reasons and rationality. He begins by noting the isomorphism between the rational transition to a psychological state from others and the derivation of a concluding proposition from premises in the deductive theoretical realm. He argues that this isomorphism led Hume to think that the rationality of the psychological transition is to be explained by the deductive validity of the derivation. Generalizing, Smith argues, Hume

concluded that the concept of a reason—that is, the concept of a consideration that justifies—must be prior to and explain the concept of rationality. The fact that there is no such isomorphism in the practical and inductive realms is therefore, Smith suggests, what led Hume to his inductive and practical skepticism. Pace Hume, however, Smith argues that we need not agree that the concept of a reason is prior to the concept of rationality. He argues that we have an independent idea of the coherence of a set of psychological states and that this is sufficient to provide us with an account of what it is for beliefs and desires to be justified. In other words, coherence provides us with the needed accounts of inductive and practical rationality, though perhaps only an account of their rationality. In the theoretical domain there are propositions to serve as objects of belief, and these propositions can be reasons for further beliefs—beliefs that can be acquired by reasoning. In the theoretical realm, then, there are not just rational transitions, but also reasons and reasoning. In the practical realm, however, there are just the rational transitions themselves: practical reasons and reasoning are figments. Furthermore, in the practical realm, perhaps there is merely means-ends rationality. But Smith concludes by asking whether practical rationality is thus restricted. He suggests that this is where the Kantians join the debate. It is, he claims, an open question whether they are right that practical coherence can be extended as far as yielding justified desires to do as morality bids.
Onora O'Neill's Kantianism, however, goes beyond mere practical coherence. She sees it as basic to Kant's thinking about practical reasoning "that reasoning can bear on action because it is formed or shaped by maxims, which have propositional structure and content." Her central concern in "Kant: Rationality as Practical Reason" (chap. 6) is to explicate Kant's account of how we could have unconditional practical reasons to do as morality requires. Unconditional practical reasons are those not based upon arbitrarily chosen ends. But then, what is their basis? Kant's proposal, O'Neill argues, is that what makes a practical reason unconditional is its universal recognizability. An unconditional practical reason is one that can be seen to be a reason for action by any rational audience—its appeal relies on no parochial concerns. Such universal appeal is captured by the categorical imperative test (O'Neill examines in detail three formulations of this): only principles of action that pass this test can be universally recognized as yielding practical reasons.
In "Duty, Rationality, and Practical Reasons" (chap. 7), David McNaughton and Piers Rawling present a view on which practical reasons are facts, such as the fact that the rubbish bin is full. This is a non-normative fact, but it is a reason for you to do something, namely take the rubbish out. McNaughton and Rawling see rationality as a matter of consistency (failing to notice that the rubbish bin is full need not be a rational failure). And they see duty as neither purely a matter of rationality nor of practical reason. On the one hand, the rational sociopath is immoral. But, on the other, morality does not require that we always act on the weightiest moral reasons: we may not be reasonably expected to know what these are. McNaughton and Rawling criticize various forms of internalism, including Williams's, and they tentatively propose a view of duty that is neither purely subjective in Prichard's (1932) sense, nor purely objective.
James Joyce's primary concern in "Bayesianism" (chap. 8) is Bayesian epistemology. Bayesianism claims to provide a unified theory of epistemic and practical rationality based on the *principle of mathematical expectation.* In its epistemic guise it requires believers to obey the *laws of probability.* In its practical guise it asks agents to maximize

their *subjective expected utility.* The five pillars of Bayesian epistemology are: (1) people have beliefs and conditional beliefs that come in varying gradations of strength; (2) a person believes a proposition strongly to the extent that she presupposes its truth in her practical and theoretical reasoning; (3) rational graded beliefs must conform to the laws of probability; (4) evidential relationships should be analyzed *subjectively* in terms of relations among a person's graded beliefs and conditional beliefs; (5) empirical learning is best modeled as probabilistic conditioning. Joyce explains each of these claims and evaluates some of the justifications that have been offered for them, including "Dutch book,"
end p.7

"decision-theoretic," and "nonpragmatic" arguments for (3) and (5). He also addresses some common objections to Bayesianism, in particular the "problem of old evidence" and the complaint that the view degenerates into an untenable *subjectivism.* The essay closes by painting a picture of Bayesianism as an "internalist" theory of reasons for action and belief that can be fruitfully augmented with "externalist" principles of practical and epistemic rationality.

In "Decision Theory and Morality" (chap. 9), James Dreier shows how the formal apparatus of decision theory is connected to some abstract issues in moral theory. He begins by explaining how to think about utility and the advice that decision theory gives us. In particular, decision theory does *not* assume or insist that all rational agents act in their own self-interest. Next he examines decision theory's contributions to social contract theory, with emphasis on David Gauthier's rationalist contractualism. Dreier's third section considers a reinterpretation of the formal theory that decision theorists use: utility might represent goodness rather than preference. His last section discusses Harsanyi's theorem.

The modeling of interaction between decision makers is the topic of Cristina Bicchieri's "Rationality and Game Theory" (chap. 10). Chess is an example of such interaction, as are firms competing for business, politicians competing for votes, jury members deciding on a verdict, animals fighting over prey, bidders competing in auctions, threats and punishments in long-term relationships, and so on. What all these situations have in common is that the outcome of the interaction depends on what the parties jointly do. Rationality assumptions are a basic ingredient of game theory, but though rational choice might be unproblematic in normative decision theory, it becomes problematic in interactive contexts, where the outcome of one's choice depends on the actions of other agents. Another basic ingredient is the idea of equilibrium play: roughly, an equilibrium is a combination of strategies, one for each player, such that each player's strategy is a best reply to the other players' choices. Thus it is individually rational for each agent to play her equilibrium strategy. But, notoriously, such individually rational play can lead to suboptimal outcomes, as in the well-known Prisoners' Dilemma. The relationship between rationality assumptions and equilibrium play is Bicchieri's main focus.

Patricia Greenspan, in "Rationality and Emotion" (chap. 11), discusses emotion as an element of practical rationality. One approach links emotion to evaluative judgment and applies some variant of the usual standards of rational belief and decision making. Fear, say, might be thought of as involving a judgment that some anticipated situation poses a

threat, and as warranted (and warranting action) to the extent that the agent has reasons for thinking that it does. In order to make sense of empathetic emotions and similar cases that do not seem to involve belief in corresponding evaluative judgments, we can modify this "judgmentalist" account by interpreting emotions as states of affect with evaluative propositional content: fear is discomfort that some situation poses a threat. If we
end p.8

also allow that the rational appropriateness of an emotional response need not be determined by the total body of evidence, in contrast to the way we assess judgments, the result is a *perspectival* account of emotional rationality. An alternative, "paradigm scenarios" approach would appeal to the causal history of an emotion as determining rationality. However, in order to assess the appropriateness of particular instances of emotion we still seem to need to refer to their propositional content or some kind of claim they make about the situation. As factors leading to action, emotions involve an element of uncontrol that is typically seen as undermining rationality but can sometimes be part of a longer-term rational strategy to the extent that states of affect modify the agent's practical options.

In "The Rationality of Being Guided by Rules" (chap. 12), Edward McClennen addresses a fundamental dilemma facing the claim that it is rational to be guided by rules. Either (1) the practical verdict issued by a rule is the same as that favored by the balance of reasons, in which case the rule is redundant or (2) the verdicts differ, in which case the rule should be abandoned. McClennen argues that we can resolve this dilemma by revising our account of practical reasoning to accord with the prescriptions of a resolute choice model. Agents in societies in which people resolutely follow, for example, a rule to keep their commitments to return favors fare better than agents in societies that lack a commitment mechanism or in which costs are incurred to enforce it.

Alfred Mele, in "Motivated Irrationality" (chap. 13), explores two of the central topics falling under this rubric: akratic action (action exhibiting so-called weakness of will or deficient self-control) and motivationally biased belief (including self-deception). Among other matters, Mele offers a resolution of Donald Davidson's worry about the explanation of irrationality: "The underlying paradox of irrationality, from which no theory can entirely escape, is this: if we explain it too well, we turn it into a concealed form of rationality; while if we assign incoherence too glibly, we merely compromise our ability to diagnose irrationality by withdrawing the background of rationality needed to justify any diagnosis at all" (1982 , 303). When agents act akratically, they act for reasons, and in central cases, they make rational judgments about what it is best to do. The rationality required for that is in place. However, to the extent to which their actions are at odds with these judgments, they act irrationally. Motivationally biased believers test hypotheses and believe on the basis of evidence. Again there is a background of rationality. But owing to the influence of motivation, they violate general standards of epistemic rationality.

In "Paradoxes of Rationality" (chap. 14), Roy Sorensen provides a panoramic view of paradoxes of theoretical and practical rationality. These puzzles are organized as apparent counterexamples to attractive principles such as the principle of charity, the transitivity of preferences, and the principle that we should maximize expected utility. The following

paradoxes are discussed: fearing fictions, the surprise test paradox, Pascal's Wager, Pollock's Ever Better wine, Newcomb's prob
end p.9

lem, the iterated Prisoners' Dilemma, Kavka's paradoxes of deterrence, backward inductions, the bottle imp, the preface paradox, Moore's problem, Buridan's ass, Condorcet's paradox of cyclical majorities, the St. Petersburg paradox, weakness of will, the Ellsberg paradox, Allais's paradox, and Peter Cave's puzzle of self-fulfilling beliefs.

2. Rationality in Specific Domains

Part 2 of this volume explores rationality's role in and relation to other domains of inquiry. It opens with chapters on rationality and psychology (chap. 15 by Richard Samuels and Stephen Stich) and rationality and gender (chap. 16 by Karen Jones). Whereas chapter 15 focuses on evidence for and against the empirical claim that we are by and large rational, chapter 16 assesses feminist challenges to what have been traditionally viewed (largely by men) as the norms that constitute what it is to be rational. In chapter 17, Carol Rovane discusses personhood and rationality. Chapter 18 is Kirk Ludwig's contribution on rationality and language. Paul Thagard's topic in chapter 19 is rationality and science. Chapter 20, by Paul Weirich, is devoted to economic rationality. Chapter 21 is Claire Finkelstein's examination of rationality and law. And in chapter 22, Peter Danielson focuses on rationality and evolution.

We will now say something in more detail about each of the chapters in part 2.

Richard Samuels and Stephen Stich, in “Rationality and Psychology” (chap. 15), explore the debate over the extent to which ordinary human reasoning and decision making is rational. One prominent cluster of views, often associated with the heuristics and biases tradition in psychology, maintains that human reasoning is, in important respects, normatively problematic or irrational. Samuels and Stich start by detailing some key experimental findings from the heuristics and biases tradition and describe a range of pessimistic claims about the rationality of ordinary people that these and related findings are sometimes taken to support. Such pessimistic interpretations of the experimental findings have not gone unchallenged, however, and one of the most sustained and influential critiques comes from evolutionary psychology. Samuels and Stich outline some of the research on reasoning that has been done by evolutionary psychologists and describe a cluster of more optimistic theses about ordinary reasoning that such psychologists defend. Although Samuels and Stich think that the most dire pronouncements made by writers in the heuristics and biases tradition are unwarranted, they also maintain that the situation is rather more pessimistic than sometimes sug
end p.10

gested by evolutionary psychologists. They conclude by defending this “middle way” and sketch a family of “dual processing” theories of reasoning which, they argue, offer some support for the moderate interpretation they advocate.

In “Rationality and Gender” (chap. 16), Karen Jones explores feminist stances toward gender and rationality. These divide into three broad camps: the “classical feminist” stance, according to which what needs to be challenged are not available norms and ideals of rationality, but rather the supposition that women are unable to meet them; the “different voice” stance, which challenges available norms of rationality as either incomplete or accorded an inflated importance; and the “strong critical” stance, which finds fault with the norms and ideals themselves. This contribution focuses on assessing the various projects—some rival, some complementary—being pursued within the third, critical camp. Jones offers a reconstruction of Catherine MacKinnon's critique of norms of rationality according to which they function to maintain relations of dominance by deauthorizing feminist claims to knowledge. Norms of rationality are thus linked to norms of credibility, and feminist rationality-critique is viewed as contributing to the naturalist project of defending norms of rationality that are appropriate for the kind of finite, embodied, socially located beings that we are.

Carol Rovane, in “Rational Persons” (chap. 17), explores eight related claims: (1) persons are not merely rational, but possess full reflective rationality; (2) there is a single overarching normative requirement that rationality places on persons, which is to achieve overall rational unity within themselves; (3) beings who possess full reflective rationality can enter into distinctively interpersonal relations, which involve efforts at rational influence from within the space of reasons; (4) a significant number of moral considerations speak in favor of defining the person as a reflective rational agent; (5) this definition of the person has led Locke and others to distinguish personal identity from animal identity; (6) although it is a platitude that a person has special reason to be concerned for its own well-being, it is not obvious how best to account for that platitude; (7) groups of human beings and parts of human beings might qualify as individual agents and, hence, as individual persons in their own right; (8) there is a sense in which the normative requirements of rationality are not categorical but merely hypothetical.

In “Rationality, Language, and the Principle of Charity” (chap. 18), Kirk Ludwig deals with the relations between language, thought, and rationality, and especially the role and status of assumptions about rationality in interpreting another's speech and assigning contents to her psychological attitudes—her beliefs, desires, intentions, and so on. The chapter is organized around three questions: (1) What is the relation between rationality and thought? (2) What is the relation between rationality and language? (3) What is the relation between thought and language? Ludwig's answers are as follows. Some large degree of rationality is required for thought. Consequently, that same degree of rationality at least is required for language, since language requires thought. Thought, however, does not require language. In answering the first question, Ludwig lays out the grounds for seeing rationality as required for thought, and he meets some recent objections on conceptual and empirical grounds. In answering questions (2) and (3), Ludwig gives particular attention to Donald Davidson's arguments for the Principle of Charity, according to which it is constitutive of speakers that they are largely rational and largely right about the world, and to Davidson's arguments for the thesis that without the power of speech we lack the power of thought.

Paul Thagard, in “Rationality and Science” (chap. 19), provides a review and assessment of central aspects of rationality in science. He deals first with the traditional question, What is the nature of the reasoning by which individual scientists accept and reject conflicting hypotheses? He also discusses the nature of practical reason in science and then turns to the question of the nature of group rationality in science. In this latter context, Thagard discusses, among other matters, his CCC (for *consensus* = *coherence* + *communication*) model, which shows how epistemic group rationality can arise in agents who communicate with each other while focusing on the explanation of observed phenomena. In the remainder of the chapter he examines whether scientists are in fact rational—that is, whether they conform to normative standards of individual and group rationality. Thagard considers various psychological and sociological factors that have been taken to undermine the rationality of science.

Paul Weirich, in “Economic Rationality” (chap. 20), examines three competing views entertained by economic theory about the instrumental rationality of decisions. The first says to maximize self-interest, the second to maximize utility, and the third to “satisfice,” that is, to adopt a satisfactory option. Critics argue that the first view is too narrow, that the second overlooks the benefits of teamwork and planning, and that the third, when carefully formulated, reduces to the second. Weirich defends a refined version of the principle to maximize utility. A broad conception of utility makes it responsive to the motives and benefits critics allege it overlooks. He discusses generalizations of utility theory to extend it to nonquantitative cases and other cases with nonstandard features.

The study of rationality as it bears on law is typically restricted to the uses made of the notion of rationality by the “law and economics movement.” Legal economists accept the traditional economic assumption that rational agents seek primarily to maximize their personal utility. What kinds of laws should a society made up of largely rational agents adopt? Legal economists supply an answer: Ideally rational legal rules, like ideally rational people, will also seek to maximize utility. They will maximize *social,* rather than individual, utility. The purpose of law, on this view, is to ensure that when individual citizens seek to maximize their individual utility, they will incidentally maximize *society's* utility. In this way, law ideally provides individual agents with incentives for efficient behavior.

Claire Finkelstein, in “Contractarian Legal Theory” (chap. 21), suggests reasons why laws that maximize social utility are not necessarily the best legal rules

for individuals that seek to maximize their personal utility. In particular, she suggests that ideally rational individuals would be unlikely to select the principle of utility maximization as the basis for choosing ideal legal rules. If Finkelstein is correct, the assumption that human beings are rational utility maximizers would have very different consequences from those legal economists have identified. Rational actor theory would be more likely to lead us to justify legal rules structured around contractarian principles—principles of agreement—than around the principle of utility maximization.

Peter Danielson's focus in “Rationality and Evolution” (chap. 22) is evolutionary game theory. Rationality and evolution are apparently quite different, applying to the acts of complex, well-informed individuals and to populations of what may be mindlessly simple

entities respectively. So it is remarkable that evolutionary game theory shows the theory of rational agents and that of populations of replicating strategies to be isomorphic. Danielson illustrates its main concepts—evolutionarily stable strategies and replicator dynamics—with simple models that apply to biological and social interactions. He distinguishes biological, economic, and generalist ways of interpreting the theory. Against the background of isomorphism, he considers three ways in which evolution and rationality differ and how two-level models may combine them. Danielson concludes with a survey of the normative significance of the unification of rationality and evolutionary game theory and some speculation about the evolution of human rationality.
end p.13

end p.14

part I
THE NATURE OF RATIONALITY

chapter 2
THEORETICAL RATIONALITY

Its Sources, Structure, and Scope
Robert Audi

The concept of rationality applies to many different kinds of things. Its widest and perhaps most complex use is in reference to persons themselves. But the concept also applies to actions, beliefs, desires, and many other elements in human life. There are, for instance, rational societies, rational plans, rational views, rational reactions, and rational emotions. A comprehensive theory of rationality must take account of this enormous diversity. [1] A full-scale account of the rationality of even one element on this list is a large undertaking and cannot be attempted here. It is possible, however, to make a brief contribution to the topic of rationality if we distinguish, as Aristotle did, between theoretical and practical rationality and concentrate mainly on one of them. In outline, the distinction centers on the contrast between the rationality of cognitions, such as beliefs, in virtue of which we are theorizing beings seeking a true picture of our world and, on the other hand, the rationality of elements, such as actions, in virtue of which we are practical beings seeking to *do* things, in particular to satisfy our needs and desires. These two dimensions of rationality are widely regarded as interconnected, and we must consider some of the relations between them, but our main focus will be on theoretical rationality.
end p.17

Belief is central for theoretical rationality. Our belief system represents the world—including the inner world of "private" experience—to us. Moreover, it is beliefs that, when true and appropriately grounded, constitute knowledge. Knowledge, in turn, is uncontroversially taken to be a "goal" of theoretical reason. Although representing theoretical reason as "seeking" a goal is metaphorical, the achievement of knowledge is widely viewed as a case of success in the exercise of theoretical reason. If, however, as skeptics have argued, our knowledge is far more limited than commonsense attributions of it would indicate, theoretical reason represents a capacity whose successful exercise is correspondingly limited.

The question of whether one or another kind of skepticism about knowledge is sustainable is large and difficult. Fortunately, it can be avoided in a brief treatment of theoretical rationality. For even if a belief does not constitute knowledge, it may be rational. I propose, then, to concentrate on conditions for the rationality of belief. [2] If these are well understood, we can account for theoretical rationality in a way that enables us to see how much of a success its exercise may be even if knowledge often eludes us. Even if skeptics are correct in claiming that our knowledge is at best highly limited, we can achieve a rational belief system whose intellectual respectability is clear.

I. Sources of Theoretically Rational Elements

A natural and promising way to begin to understand rationality is to view it in relation to its sources. The very same sources yield justification, which is closely related to rationality. These sources are also central for reasonableness, which implies rationality but is a stronger notion. Our reasonable beliefs, like our justified ones, are rational, but a belief that is rational—at least in the minimal sense that it is not irrational—may be (beyond avoiding inconsistency and other clear defects) simply plausible to one, sometimes in the way a sheer speculation often is, and may fail to be justified or reasonable, as one may later admit. At times I will connect rationality with these concepts, but to avoid undue complexity I will focus chiefly on theoretical rationality, with rational belief as the central case.

The Classical Basic Sources of Rationality

If, in the history of epistemology, any sources of the rationality of belief deserve to be called the classical basic sources, the best candidates are perception, memory,

end p.18

consciousness (sometimes called *introspection*), and reason (sometimes called *intuition*). Some writers have shortened the list under the heading "experience and reason." [3] This heading is apt insofar as it suggests that there is some unity among the first three sources and indeed the possibility of other experiential sources of rational belief; it is misleading insofar as it suggests that experience plays no role in the operation of reason as a source of rational belief (and of justification and knowledge). Any operation of reason that occurs in consciousness—for instance, engaging in reasoning—may be considered a kind of intellectual experience. The reflection or other exercise of understanding required for "reason" to serve as a source of rational belief is certainly one kind of experience.

Let us first explore what it is for a source to be basic and some of the conditions under which beliefs it yields are rational. We can then consider what kind of source might be

nonbasic and whether the four standard basic sources are the only basic sources of theoretical rationality.

I take it that a *source* of (theoretical) rationality (or justification) is roughly something in the life of the person in question—such as perception or reflection—that characteristically yields rational beliefs. I also take it that to call a source of theoretical rationality (or of justification) *basic* is to make a comparative statement. It is not to rule out every kind of dependence on *anything* else, but simply to say that the source yields rational belief without positive dependence on the operation of some other source of rationality (or of justification). We might begin with perception.

Perception

On the basis of perception, I might rationally believe and indeed know that the clock says ten; I know this by virtue of seeing its face displaying that time. On the basis of brief reflection, I might rationally believe (and know) that if one proposition entails a second and the second entails a third, then if the third is false, so is the first. To be sure, this belief is not possible without my having the concepts required to understand what I believe, but that conceptual requirement is not a positive dependence on a source of rationality.

It may seem that the perceptual belief can be rational only if I remember how to read a clock and that therefore perception cannot yield rational belief independently of memory, which is also a source of rational beliefs. It is true that rational perceptual belief may depend on memory in a certain way. But consider this. A being could acquire the concepts needed for reading a clock at the very time of seeing one, and hence would not need to *remember* anything in order to form the belief (at that very time) that the clock says ten. One possibility is the creation of a duplicate of someone like me: reading a clock would be possible at

end p.19

his first moment of creation. It appears, then, that although perceptually grounded rational belief ordinarily depends in a certain way on memory, neither the *concept* of perception nor that of rational perceptual belief (or perceptual knowledge) is *historical.* That of memory, however, is historical, at least in this sense: one cannot remember something unless one has *retained* it in memory over some period of time.

One might think that perception is not a basic source of theoretical rationality because of the way it depends on consciousness. The idea would be that one cannot perceive without being conscious; hence, perception cannot yield rational belief (or knowledge) apart from the operation of another source of it. Let us grant that perception requires consciousness. [4] If it does, that is not because consciousness is a precondition or a causal requirement for perception, but because perception *is* a kind of consciousness: consciousness of an external object. The dependence would be constitutive rather than operational. We might then simply grant that perception is perceptual consciousness and treat only "internal consciousness" (consciousness of what is internal to the mind) as a source of rational belief (or of knowledge) distinct from perception.

Internal consciousness, understood strictly, occurs only where its object is either internal in the way images and thoughts are (roughly, phenomenal) or abstract, as in the case of concepts and (presumably) numbers. On a wider interpretation, internal consciousness

might have dispositional mental states, such as beliefs, desires, and emotions, among its objects. But even when this occurs, it seems to be *through* consciousness of their manifestations that we are conscious of such states, as where we are conscious of anxiety through being aware of unpleasant thoughts of failure.

Philosophers in the sense-datum tradition have held that ordinary perception of physical objects is also in a sense indirect, being "through" acquaintance with "ideas" of them that represent them to us. But an account of theoretical rationality need not be committed to such a *representationalism.* One can plausibly hold both that perception requires a sensory experience and that external objects are directly perceived—and that in that sense we are directly conscious of them, as opposed to being conscious of some interior object that represents them to us. [5]

To be sure, one might also treat consciousness as a kind of perception: external perception where the perceived object is outside the mind, internal where that is inside. But abstract objects are not "in" the mind, at least in the way thoughts and sensations are. In any case, it is preferable not to consider consciousness of abstract objects as a kind of perception. One reason for this is that there is apparently a causal relation between the object of perception and whatever sensation or other mental element constitutes a perceptual response to it, and it is at least not clear that abstract entities have causal power, or at any rate the requisite kind. [6] This issue is too large to pursue here, but it may be enough to

end p.20

note that not all mental phenomena seem to be either perceptual in any sense or to be directed toward abstract objects. Consider, on the "passive" side of mental life, an idle daydream, or, on the "active" side, planning. Neither need concern the abstract, nor must we suppose that there are objects in the mind having properties in their own right. [7] It would be unwise to assume that perception exhausts the activity of consciousness.

It does appear, however, that we may take the concept of perception to be a partly causal notion. If you see, hear, touch, taste, or smell something, then it affects you in some way. And if you may be said to perceive your own heartbeat or even your own anxiety, this is owing to their causing you to have some experiential impression analogous to a sense-impression you might have through the five senses. Conceived in this way, perception is not a *closed concept:* it leaves room for hitherto unfamiliar kinds of experiential response to count as the mental side—the subjective response side, one might say—of perceiving an object and indeed for new or unusual kinds of objects to be perceptible. [8] This is not the place, however, to give an account of exactly what perception is. My point is that there may be perceptual sources of theoretical rationality other than the familiar ones. The concept of theoretical rationality is surely no more closed than is the notion of a perceptual source of belief.

Memory

If, in speaking of perception, we are talking about a capacity to perceive, in speaking of memory we are talking about a capacity to remember. But remembering, in the sense of having a veridical memory of something, does not exhaust the operation of our memorial capacity to the extent that perceiving, in the sense of having a veridical perception of something, exhausts the operation of the perceptual capacity. There is also *recalling,* which entails but is not entailed by remembering; *recollecting,* which is similar to recalling but tends to imply an episode of (sometimes effortful) recall, usually of a sequence or a set of details and often involving imagery; and memory *belief,* which may be mistaken and does not entail either remembering or even recalling. It seems, however, that remembering that *p* (where *p* is some arbitrarily chosen proposition) entails knowing it; and we also speak of knowing things from memory. When we do know things (wholly) in this way, it is not on the basis of other things we know. One may know a theorem from memory *and* on the basis of a simple proof from an axiom. But where one knows *p* wholly from memory—simply by virtue of remembering it—one does not at the time know it on the basis of knowing or believing anything else.

These points make it natural to think of memory as a basic source of knowledge as well as of rational beliefs that fall short of knowledge (say because they are false or based on too weak a memory impression). But I doubt that memory is a basic source of knowledge. It *is* an epistemically *essential source;* that is, what we think of as "our knowledge," in an overall sense, would collapse if memory did not sustain it: we could know only what we could hold in consciousness at the time (at least this is so *if* what we know dispositionally at a time must be conceived as held in memory at that time, even though it is true then that if we *were* to try to bring any one of the propositions to consciousness then, we would normally have it there then). [9] By virtue of playing this role, memory is an epistemic source in an important sense. But surely one cannot know anything from memory without coming to know it through some *other* source. If we remember it and thereby know it, we *knew* it; and we must have come to know it through, say, perception or reasoning. [10]

If, however, memory is not a basic source of knowledge, it surely *is* a basic source of theoretical rationality (and of justification for belief). Just how it plays this role is not easy to capture. But consider believing that last week one telephoned a friend. There is a way this belief—or at least its propositional object—can present itself to one that confers some degree of justification on the belief (I think it can confer enough to allow the belief to constitute knowledge if one is correct and there is no defeater of one's would-be knowledge, but there is no need to try to show that here). Someone might object that it is only by virtue of knowledge, through consciousness, of one's memorial images that we can be justified in such beliefs, but I very much doubt this. [11] A remembered proposition can surface in consciousness without the help of images, and often spontaneously, upon the need for the proposition in answer to a question about the relevant subject or as a premise for an argument that one can see to be needed to justify a claim one has made. In the light of the points made about memory so far, I suggest that it is an *essentialsource of knowledge* and a *basicsourceof justification.* In the former case it is *preservative,* retaining knowledge already gained; in the latter it is *generative,* producing justification not otherwise acquired. Given the way that memory can preserve belief and indeed knowledge—retaining them even when any premises we may have initially had as a basis

for them are forgotten—it has another positive epistemic capacity. It can be a preservative source of basic knowledge even without being a basic source of it. Knowledge from memory need not be based (inferentially) on other knowledge or belief and hence can be basic; but since the knowledge must be acquired through another source, memory depends operationally on that source and is not a basic source of the basic knowledge in question. Memory can, then, produce knowledge that is basic in the order of knowings even though memory itself is not basic in the order of sources.
end p.22

Consciousness

Consciousness has already been mentioned as a basic source of rational belief (and of knowledge). It seems clear that if any kind of experience can yield rational belief, it is introspective consciousness of what is presently in one's mind. Even philosophers who take pains to give skepticism its due, such as David Hume, do not deny that we have knowledge—presumably noninferential knowledge—of our own current mental life. [12] Suppose those who deny direct realism—roughly the view that we perceive external objects without the mediation of objects constituting mental representations of them—are right and some form of representative realism (the mediation view just sketched) is true. Then it is only consciousness of the inner world—or at least of whatever can exist "in" consciousness—that is a basic perceptual source, since outer perception (consciousness of the external world) is not a basic source. But the inner world is a very important realm. It might include abstract objects, such as numbers and concepts, as well as sensations, thoughts, and other mental entities. (This would not imply that abstract objects are mental; the sense in which they are in the inner world is a matter of their direct accessibility to thought, not of their mode of existence.) And for nonskeptics, even if we do not directly perceive external objects, we may still have knowledge of them through perceptual experience that, like experience of sense-data, represents them.

Reason

When we come to reason, there is, as with memory, a need to clarify what aspects of this general capacity concern us. Like "memory," the term "reason" can designate quite different things. One is reflection, another reasoning, another understanding, and still another intuition. We reflect on a subject, reason from a hypothesis to see what it implies, understand a concept or proposition (sometimes only after reflection), and intuit certain truths. These are only examples, and there is overlap. Any of the objects in question must be understood (adequately, though not perfectly) if it is to be an object of reason, and understanding the truth of a proposition—say that *p*—that one intuits may require reflecting on it: understanding may not come quickly or even easily.

It will help to focus on a simple example. Consider the logical truth that if all human beings are vulnerable and all vulnerable beings need protection, then all human beings need protection. We can reason from the "premises" (expressed in the *if* clause) to the "conclusion" (expressed in the *then* clause), but an assertive use of this conditional sentence need not represent giving an argument. Moreover, the proposition it expresses is not the kind that would (normally) be known by

end p.23

reasoning. It would normally be rationally believed (and known) by "intuition" or, in the case in which such direct apprehension of the truth does not readily come to a person, by reflection that indirectly yields understanding. (The *conclusion*—that all human beings need protection—may of course be known wholly on the basis of reasoning from the premises. One's knowledge of it then depends on one's knowledge of them, and that knowledge surely requires reliance on a different basic source. But the proposition in question is the conditional one connecting the premises with the conclusion, and knowledge of *that* does not require knowledge of either the former or the latter.)

Reasoning

It turns out, then, that "reasoning" is not a good term for the ratiocinative basic source we are considering. Indeed, if we distinguish reasoning from reflection of a kind that yields knowledge that *papart from* reliance on independent premises, it is best not to use "reasoning" in describing this source. What seems fundamental about the source is that when knowledge of, or justification for believing, a proposition comes from it, it derives from an exercise of reason regarding the proposition. This may take no time beyond that required to understand a sentence expressing the proposition (which may be virtually none; nor need we assume that all consideration of propositions is linguistically mediated, as opposed to conceptual in some sense). Here it is natural to speak of intuiting. But the proposition may not be so easily understood, as (for some people) is the case with the proposition that if *p* entails *q* and *q* entails *r,* and either not-*q* or not-*r* is the case, then it is false that *p.* In this case it is more natural to speak of reflection. In either case the source seems to operate by yielding an adequate degree of understanding of the proposition in question and thereby knowledge. It does not appear to depend (positively) on any other source and is plausibly considered basic. [13]

It also seems clear that reason is a basic source of rational belief (as of justification and knowledge). Such simple logical truths as those with the form of "If all *As* are *Bs* and all *Bs* are *Cs,* then all *As* are *Cs*" can be both justifiedly believed and known simply on the basis of (adequately) understanding them. In at least the vast majority of the cases in which reason yields knowledge, it apparently also yields justification. It can, however, yield justification for a belief without grounding that belief in a way that renders it a case of knowledge. This may occur even where the belief is true. [14]

The more common kinds of justified beliefs that do not constitute knowledge are *not* true. Careful reflection can make a proposition seem highly plausible even where later reflection shows it to be false. If we are talking only of prima facie

end p.24

(hence defeasible) rationality (and justification), there are many examples in logic and mathematics. Consider Russell's paradox. [15] There seems to be a class of non-teaspoons

in addition to a class of teaspoons. The latter class, however, is plainly not a teaspoon, since it is a class. So, it is a non-teaspoon and hence a member of itself. The same holds for the class of non-philosophers: being a non-philosopher, it is a member of itself. It now seems that there must be a class of such classes—a class of all and only those classes that are not members of themselves. But there *cannot* be one: this class would be a member of itself if and only if it is *not* a member of itself. Thus, what appears, on the basis of an exercise of reason, to be true may be false.

It may be objected that it is only inferentially that one could here believe there is a class of all and only classes that are not members of themselves and that therefore it is not only on the basis of the operation of reason that one would believe this. But surely we may take reasoning to be *one* kind of such operation, particularly deductive reasoning. It is true that the *basic kind* of knowledge or justification yielded by a source of either is noninferential; there is no good reason, however, to rule that inferences may not be included among operations of reason.

To be sure, there is still the question whether inference depends on the operation of memory, in the sense that one may draw an inference from a proposition only if one *remembers* it. This seems false. One can hold some simple premises before one's mind and at that very time draw an inference from them. People vary in the relevant *inferential memorial capacities,* as we might call them. If we allow that rationality (or knowledge or justification) deriving from simple inferences such as those in question here need not depend on memory, we may conclude that it can be on the basis of inferential reason that the proposition in question is rationally believed. It is a contingent matter whether such an inference *does* depend on the operation of memory. If one must write down the premises to keep track of them, it would (unless visual or other sensory representation of them enabled one to keep them in mind as one draws the inference). If, however, one can entertain the premises and conclusion together and at that time see their logical relation, it does not. The distinction between these two cases is not sharp, but it is often quite clear.

Fallibility and Defeasibility

Even reason should not be considered an *infallible source* of rationality (or of justification or knowledge): one whose every cognitive deliverance is true. One could think too superficially where one should know better, or infer a conclusion that obviously does not follow. In many such cases one might form a false belief. One might also form a belief that is not even rational (though it need not be

patently irrational either—I am thinking of cases of sloppiness or inattention that occur without blatant offense against reason). To call a source basic is to affirm a measure of epistemic autonomy; it is not to give a wholesale epistemic guarantee. It is perhaps not obvious that every cognitive “deliverance” of a basic source has prima facie rationality (or prima facie justification); but this is a plausible view, if we (1) take a cognitive deliverance of a source to be a belief *based* on it and not merely caused by it, and (2)

allow that a belief can be prima facie rational even when its rationality is massively overridden. Let us assume (1) and (2). Plainly this would not entail indefeasible rationality (or indefeasible justification). If we suppose, then, that there would be no rational belief (or knowledge or justification) *without* basic sources of it, we still cannot reasonably conclude that every belief those sources deliver is rational on balance (or justified on balance or, if true, constitutes knowledge.)

To be sure, even simple logical truths can be rationally believed (or known) on the basis of testimony, as where someone who is logically slow first comes to know one through the testimony of a teacher. Here the immediate basis of the belief, the testimony, is empirical. But can such truths be known or justifiedly believed without someone's depending on reason *somewhere* along the line? It would seem that the teacher must depend on it, or on testimony from someone who does, or who at least must rely on testimony from someone else who depends on reason, and so forth until we reach a person who knows it a priori. [17] Knowledge through testimony, then, even if direct in the sense of "noninferential," might be called *secondary,* in contrast with the kind that does not depend (in the way testimony-based knowledge characteristically does) on any other knowledge (or justified belief) and is in that sense *primary.*

Might we, however, make the parallel claim for perceptual and introspective cases? Could anyone, say, know the colors and feel of things if no one had perceptual knowledge? If we assume the possibility of an omnipotent and omniscient God, we might have to grant that God could know this sort of thing by virtue of (fully) knowing God's creation of things with these colors and textures. Still, wouldn't even God have to know what these properties are *like* in order to create the things in question with full knowledge of the nature of the things thus created? Suppose so. That knowledge is arguably of a phenomenal kind; if it is, the point would show only that for a certain kind of knowledge consciousness is a *uniquesource:* the only kind capable of delivering it. Perhaps it is unique, and perhaps the same holds for rational beliefs of the kinds of propositions in question. If reason and consciousness are not only basic, but also the only unique sources, one can understand why both figure so crucially in the epistemology of Descartes, or indeed any philosopher for whom what is accessible to conscious experience and to thought is epistemically fundamental in the far-reaching way that is implied by the combination of basicality and uniqueness.

Testimony

The four standard basic sources do not include testimony. But I have indicated an important epistemological role for it. It is rightly taken to be a source of a great many of our rational beliefs. In human life as we know it, testimony (in the broad sense of people saying things to us) is essential for the rationality of a vast proportion of our beliefs about the world. It is not, however, a basic source of theoretical rationality. For one thing, it can yield justified belief (or knowledge) in the recipient only if perceived by that person, say heard or read. The basic sources, by contrast, operate autonomously in their respective realms. There is much to say about just how testimony figures in grounding theoretical rationality. To say that it is not basic is to describe how it operates; it is not at all to

diminish the scope or importance of its role. It is time, however, to consider a different kind of source.

II. Coherence

An alternative to the position developed so far is that a major source of theoretical rationality, and perhaps *the* basic source of it—particularly in the form of justification for belief—is coherence among one's beliefs. Consider my belief that the home team has won a football game, based on hearing revelers at the time the game was to end. Isn't my belief that they have won justified by its *coherence* with the beliefs that people noisily celebrate football victories, that there is no other explanation of the celebratory noises, that I have noticed such a pattern before in cases of victory? And suppose I lose justification, owing to undermining evidence, as where I suddenly see a wedding party. Isn't the justification of my belief that the home team won undermined mainly by its *in*coherence with my present beliefs that the noise is from a wedding party? Let us explore the role of coherence in justification.

Coherence, Incoherence, and Noncoherence

It is difficult to say what constitutes coherence. The notion is elusive, and there are highly varying accounts. [18] But this much is clear: we cannot assess the role of coherence in justification unless we distinguish the thesis that coherence is a basic
end p.27

source of justification from the thesis that *in*coherence can defeat justification. The power to defeat is destructive; the power to provide grounds is constructive. To see that the destructive power of incoherence does not imply that coherence has any basic constructive power, we should first note that incoherence is not the contradictory of coherence, its mere absence. It is something with a definite negative character: two beliefs that are logically and semantically irrelevant to each other, such as my beliefs that the sun is shining and that I am thinking about sources of knowledge, are neither mutually coherent nor mutually incoherent. They are simply noncoherent. The paradigm of incoherence is blatant logical inconsistency; positive coherence is widely taken to be far more than mutual consistency, yet far less than mutual entailment.

Clearly, that incoherence can defeat justification does not imply that coherence can create it. If it does create it (which is far from obvious), seeing this point is complicated because wherever coherence is plausibly invoked as a source of justification, one or more of the four standard sources apparently operates in a way that provides for an explanation according to which *both* the coherence and the justification arise from the same elements responsible for we might call the *well-groundedness* of the belief in question. [19] This is best seen through cases.

Consider my belief that a siren is sounding, grounded in hearing the distinctive shrill crescendo. This appears to be justified by the relevant auditory impressions, together with background information about what the corresponding sounds indicate. If, however, I acquired a justified belief that someone is imitatively creating the blare, my justification for believing that a siren is sounding would be undermined by the incoherence now in my belief system. Does the defeating power of incoherence imply that my original justification requires coherence among my beliefs, including the belief that no one is doing that? Does one even *have* that belief in such a case? It would surely not be normal to have it when there is no occasion to suspect such a thing. But suppose the belief were required. Notice how many beliefs one would need in order to achieve coherence of sufficient magnitude to be even a plausible candidate to generate the justification in question, for example that my hearing is normal, that there is no other machine nearby that makes the same grating sounds—it is not quite clear how far this must go. Do we even form that many beliefs in the normal cases in which we acquire justified beliefs of the ordinary kind in question? To think so is to fall victim to a kind of intellectualism about the mind that has afflicted coherentist theories and opposing accounts of justification alike. [20]

A further analogy may help to show how incoherence can be a defeater of justification without (1) its absence, (2) beliefs that it is absent, or (3) justification for believing something to this effect being a source of justification. One's job may be the source of one's income, yet vulnerable to severe economic depression, since that might eliminate the job. It does not follow that the absence of a depression is a source of one's income. Surely it is not. Even positive economic

conditions are not a source, though one's source of income *depends* on them. The idea of (positive) dependence is central in understanding that of a source.

It must be granted that there is a negative sense in which one's job does depend on the absence of a depression; but that dependence—a kind of vulnerability—is too negative a condition to count as a source (much less a ground) of income. Even a good economy does not give one an income. Nor does it explain why one has the income. Similarly, we might say that one's justification negatively depends on the absence of defeaters and positively depends on one's sources. But negative dependence on incoherence does not imply positive dependence on anything in particular, including coherence, as a source, any more than an income's negative dependence on the absence of a depression implies any particular source of that income.

Epistemic Enablers versus Epistemic Grounds

Nothing can serve as a source of anything without the existence of indefinitely many *enabling conditions.* Some of these are conceptual. One may, for instance, be unable to believe a proposition even when evidence for it is before one. If a child has no concept of a flight recorder, then seeing one removed from the wreckage of an airplane will not

function as a source of justification for the proposition that it was recovered. Other enabling conditions are psychological, concerning our dispositions relevant to forming beliefs. If my sensory receptors are malfunctioning, or if I do not respond to their deliverances by forming beliefs in the normal way, then I may fail to be justified in certain perceptual beliefs. In this way, *contextual* variables are crucial for determining whether a belief is rational (or justified) in a given case; but that point is one that both well-groundedness (and in that sense "foundationalist") views and coherentist views can accommodate.

Specifying a source provides both a genetic explanation of where a thing comes from and, through supplying a ground, a contemporaneous explanation of why it is as it is; enabling conditions, by contrast, provide neither. Taken together, they explain its possibility, but not its genesis or its character. It is neither correct nor theoretically illuminating to construe the absence of the enabling conditions as part of the source or as a ground. They are indispensable, but their role should be understood in terms of the theory of defeasibility rather than the theory of sources or of positive grounds.

The importance of incoherence as a defeater of justification, then, is not a good reason to take coherence to be a source of justification. This by no means implies that justification has no important relation to coherence. Indeed, at least normally, justified beliefs will cohere, in one or another intuitive sense, with other

end p.29

beliefs one has, typically other justified beliefs. Certainly, wherever there is justification for believing something, there at least tends to be justification for believing a number of related propositions and indeed for believing a coherent set of them. This is easily seen by reflecting on the point that a single perceptual experience provides information sufficient to justify many beliefs: that there is a street before me, that someone is tooting horns on it, that this charivari is louder than my radio, and far more.

The conception of sources of rational belief (and of knowledge and justification) that I have sketched provides a way to explain why coherence apparently accompanies rational and justified beliefs—actual and hypothetical—namely that rationality and justification are ultimately grounded in the same basic sources. In sufficiently rich forms, coherence may, for all I have said, commonly be a *mark* of rationality and justification: an indication of their presence. The coherence conception of rationality and justification, however, does not well explain why they apparently depend on the standard sources. Indeed, as an internal relation among beliefs, coherence may be as easily imagined in artificial situations where the coherence of beliefs is unconstrained by our natural tendencies. In principle, wishful thinking could yield as coherent a network of beliefs as the most studious appraisal of evidence. [21]

Conceptual Coherentism

One kind of coherence, to be sure, is entirely consistent with the well-groundedness conception of theoretical rationality that goes with taking it to derive from basic sources

in the ways I have suggested. To see this, note first that one cannot believe a proposition without having the concepts that figure essentially in it. Whereof one cannot understand, thereof one cannot believe. Moreover, concepts come, and work, in families. They do not operate atomistically. This point is the core of a coherence theory of conceptual function: of the acquisition of concepts and their operation, most notably in discourse, judgment, and inference. That theory—*conceptual coherentism,* for short—is both plausible and readily combined with the view presented here. For instance, I cannot believe, hence cannot rationally believe, that a siren is sounding unless I have concepts of a siren and of sounding. I cannot have these unless I have many other concepts, such as those of signaling, hearing, and responding. Granted, no one highly specific concept need be necessary, and various alternative sets will do. In part, to have a concept (of something perceptible) is (at least for remotely normal persons) to be disposed to form beliefs under appropriate sensory stimulations, say to believe a specimen of the thing to be present when one can see it and is asked if there is such a thing nearby. Thus, again it is to be expected that from a single perceptual experience, many connected propositions will be justified for the perceiver.
end p.30

The coherence theory of conceptual function belongs more to semantics and philosophy of mind than to epistemology. But it has profound epistemological implications. That concepts are acquired in mutual relationships may imply that rationality and justification do not arise atomistically, in one isolated belief (or desire or intention) at a time. In that minimal way, they may be "theory-laden"—though the term is misleading in suggesting that having a family of concepts entails having a theory. None of this implies, however, that once a person acquires the conceptual capacity needed to achieve justification, justification cannot derive from one source at a time (nor need we suppose that concept-formation develops earlier than, or in isolation from, the formation of rational belief). This theory of conceptual acquisition and competence is also quite consistent with the view that, far from deriving from coherence, justification, by virtue of the way it is grounded in its sources, brings coherence with it.

III. Theoretical Rationality and the Structure of Cognition

We have seen what sorts of bases ground the rationality of beliefs and, often, justification and knowledge as well. But a person does not achieve theoretical rationality simply by having beliefs properly grounded in one or even all of the basic sources. Those beliefs are, as we saw, noninferential. If we never formed beliefs on the basis of those, it would be as if we laid only the foundations of a building and never erected even a single story upon them. Even if one could survive simply on the ground, there is much that cannot be seen without ascending to higher levels. Some things we cannot know or even rationally believe except by inference (or through a similar building process) from what we believe through the basic sources. Perception alone, for instance, yields no theories, and intuition

unaided by inference, even if it provides premises for the branches of mathematics, does not automatically yield any theorems.

Inference and Inferential Grounding

It is largely because inference is so pervasive in our lives as rational beings that *reasoning* is considered so important for our rationality. For inference is the central case of reasoning and, if the latter term is used strictly and contrasted with "thinking," arguably the only case. I have already suggested that no process of reasoning is required for a belief to be based, in an inferential way, on one or more others; but in fact it would be at best abnormal for any of us *never* to do reasoning, conceived roughly as passing, under the guidance of an appropriate principle, from considering one or more propositions ("premises") to another (the "conclusion"). We cannot say "from at least one *believed* proposition," because there are inferences we make simply to see what follows from something—sometimes with a view to refuting it by deriving a contradiction. Here we may make a *non-belief-forming inference:* we infer the contradiction only to reject it—and indeed thereby infer (and believe) the negation of the proposition being tested. And we cannot say that the person must *believe* the appropriate principle, since one may be guided by a principle one is just trying out or, as is common with children learning to reason, one may be guided by a principle one cannot formulate and before one has internalized it in the way required for believing it.

There is no precise limit to the number of beliefs that can be inferentially grounded on beliefs that are "basic" in the sense of "noninferential," and no limit to the length of a chain of inferences. One can infer conclusions from one's conclusions, further conclusions from them, and so forth. Our rationality is not directly proportional to the number of beliefs we have, nor even to the sheer quantity of our rational beliefs or knowledge. Some rational beliefs and knowledge are trivial, say that there is more than one speck of dust in this room. Moreover, a person who is theoretically rational must have a belief system with certain structural features. Let me describe these in outline.

I have already indicated that some degree of coherence among beliefs is to be expected in rational persons. We may add that other things equal, a more coherent set of beliefs tends to be more rational overall and to bespeak greater rationality in their possessor than a less coherent set. But there is a further point of major importance. There is a sense in which rational beliefs must cohere *with experience.* If I am visually experiencing black printing on white paper, I should (normally) believe that such print is before me, at least if I consider whether it is; and I (normally) must not believe that I am seeing red print. If thunder rattles the windows, I should normally believe they are rattling, or something to that effect. Experience of the inner world is similarly a basis with which rational beliefs must normally cohere. If I am silently reciting some lines of poetry, then (at least if I consider the matter) I should normally believe that I am silently reciting some lines.

Some Modes of Belief-Formation

A more general way to put the point is to say that belief-formation and indeed belief-retention should be adequately responsive to experience. This does not re
end p.32

quire that in the course of ordinary experiences we form the vast numbers of beliefs we can form, say at least one for all the truths about a room that are in some sense perceptually represented to us upon entering it—that the sofa is blue, that there are three scatter rugs, that the straight chair is at least a foot taller than the sofa, that the carpeting is seamless. [22] But we must be *disposed* to form beliefs of propositions that our present experience makes evident to us and *not* to form beliefs of obvious contraries of those propositions.

The kind of responsiveness to experience I am describing may be viewed as a kind of coherence; but if it is so viewed, we must not conclude that its importance supports epistemological coherentism, conceived as roughly the view that the *basis* of cognitive rationality and cognitive justification is coherence among beliefs. That rational beliefs must in general cohere with experience, far from implying that their mutual coherence produces rationality, expresses a constraint on the kinds of beliefs whose mutual coherence is a reason to expect them to be rational. [23] For if none of our beliefs is grounded in experience—including the kind of reflective experience that yields beliefs of self-evident propositions—then any coherent set might be considered rational, including one that is internally coherent but inconsistent with what is supported by the person's experience, as in typical cases in which a mental illness leads to an elaborate system of delusions.

Foundationalism

The grounding role that experience plays in determining theoretical rationality is central for foundationalist theories of that notion. A moderate kind of foundationalist theory of rationality that seems highly plausible says that if there are any rational beliefs at all, there are some that are noninferential, and that any other rational beliefs derive enough of their justification from support they receive from one or more foundational beliefs so that if (other things remaining equal) they lost any support they have from other sources, they would remain rational. By contrast, a moderate coherentist theory of rationality would deny that noninferential rationality is needed and would give to coherence among beliefs the same importance foundationalism gives to their experiential grounding. This is not the place to compare and contrast the two theories in detail; I am here suggesting that for some of the reasons indicated above, a moderate foundationalist approach provides a more plausible account of theoretical rationality. Such an approach is compatible, it should be added, with reliabilism, virtue epistemology, contextualism, and other plausible epistemological perspectives. [24]

If theoretical rationality requires a certain kind of responsiveness to experience, and if the beliefs that are direct (noninferential) responses to it are basic in one's cognitive structure, then our belief system should be expected to have certain
end p.33

psychological features. Some of our beliefs should be noninferential and others based on them. Many may be based on a single one; many basic ones may support a single belief. There is no precise limit here. Nor is there any precise limit to how many links there can be between a basic element and elements based on it.

Belief-Change

One's system of beliefs, may, moreover, change greatly over time. A belief that is noninferential at one time may be inferential later, when one has acquired a premise for it. A belief inferentially based on premises may be retained in memory long after the premises are forgotten and hence be noninferential—memorially direct, we might say. Where the memory impression grounding the belief meets certain conditions (say, is steadfast and not in conflict with any other impression or belief one has), retention of the belief may be rational. Here both a kind of coherence and a connection with foundational elements is pertinent. For instance, if the belief is the kind I can rationally suppose I acquired from adequate evidence, as with a strong memory belief that a certain novel is by Balzac, I have no need for a premise. Retaining the belief coheres with what I (rationally) believe about my evidence base, and memory impressions themselves play a positive role in grounding the rationality of beliefs.

Since I am leaving skepticism aside, I am assuming that our rational beliefs, whether basic or not, can be an adequate ground for either inductive or deductive extension. We can acquire new rational beliefs—for instance, by inference to the best explanation, as where we come to believe that a train is late because that best explains why a visiting speaker is late for the seminar. We can also acquire them by deduction, as where we infer theorems from axioms. To be sure, one can be rational in holding a belief at one confidence level but not at a higher one. I have been for the most part ignoring this variable, as well as the related notion of *degrees of belief;* but this notion can be accounted for using the raw materials we have been considering. [25] Other things equal, the better one's grounds for *p,* the greater the confidence one may rationally have toward it.

Plainly, extension of our rational belief system may also occur as a result of testimony. There is some controversy over whether the resulting beliefs are genuinely noninferential. [26] I do not see that they need to be; but in any case, since their source is (in my judgment) not basic, they are best conceived as instances of extension beyond the beliefs that arise as responses to experience in the realm of the basic sources. An immensely wide and indefinite variety of rational beliefs may arise from testimony. Not just any testimony is credible, of course; but perhaps we might say that normally, we may rationally believe

what people attest to unless we have reason to doubt it. It is a contingent matter how often that is in a given person's social experience.
end p.34

IV. The Scope of Theoretical Rationality

We have now seen what sorts of grounds, basic and inferential, theoretically rational elements have, and what kind of structure a system of rational elements has in a rational person. So far, however, the scope of theoretical rationality has been left largely open. Are there propositions, such as simple logical truths, that any rational person must believe? And are there limits to the range of propositions that can be objects of rational belief in persons like us? (I assume that omniscience is not possible for finite minds like ours.) Let me address these in order.

Beliefs versus Dispositions to Believe

I have already noted that being guided by a logical principle can apparently precede the believing of it. Moreover, there are propositions of many kinds that a normal rational person will believe upon considering them, say (for readers of this page) that there are more than 103 letters written here. But although our potential for forming beliefs is incalculably wide, we are highly limited in what propositions, particularly logical truths and elementary propositions made obvious by our experience, we can *disbelieve.* Nonetheless, even if this requirement carries with it a strong *disposition to believe* the negations of those propositions, actual belief of the latter is not a condition for rationality. On the overall view I have been stressing, theoretical rationality is above all a kind of responsiveness to grounds (the kind in virtue of which cognitions are justified). In the basic cases, it is responsiveness to experiences, in particular to experiential *grounds;* in the other instances it is above all responsiveness to *beliefs* formed on the basis of experience (the second case is typically one of inferential responsiveness). The basic cases of responsiveness to experience apparently do not require believing any particular propositions.
Indeed, it appears that the experiential responsiveness central for rationality does not even entail *having* beliefs, as opposed to dispositions to form them, at all. The brain could be manipulated in such a way that for a short time one is left with no beliefs, but only capacities and dispositions to form beliefs. It is not clear what consciousness would be like at such a moment; but a model for understanding it might be an exercise in which, perhaps with the help of skeptical reflection, one suspends judgment on a plausible proposition one is considering. There may be a limit to which this ability can be developed in a rational person, but perhaps with the aid of skillful brain manipulation nonbelief could be induced relative to all of the propositions in one's belief system.
end p.35

Whatever we say about the question whether a theoretically rational person must have beliefs, and indeed some that are theoretically rational, it is plain that the central question here concerns what is required for appropriate responsiveness (direct or indirect) to experience. If that is possible for a person having no beliefs, but instead only suitable capacities and dispositions to form beliefs, then a rational person need not have beliefs.

Some Limitations on Rational Belief

Our second question about the scope of theoretical rationality is even more difficult. It might seem that we could say that the scope of theoretically rational elements possible for us is limited only by our finitude. After all, isn't it possible that an omnipotent God may simply endow one with a rational belief of any proposition that, given one's finite capacity for understanding, is comprehensible to one? This is not unconditionally so (at least on the plausible assumption that divine power operates within the domain of the logically possible). We would not rationally believe a proposition simply because God had implanted the belief in us or even because it is an a priori truth. Rational belief (and indeed rational cognition of any kind) requires adequate grounds, not just causation by a perfect being or eminently credible true content. It turns out, I suggest, that the limits of our rational beliefs extend no further than our rationality-conferring grounds.

Given this dependence of rationality on grounds, the scope of theoretical rationality for one person will be quite different from its scope for another. Each of us has different experiences, and people differ widely in inferential powers. An intellectually normal person, however, must have both a minimal responsiveness to experience—including the intellectual experience of reflection on simple a priori matters—and minimal logical powers of inference. This has been illustrated with respect to sensory experience and the consideration of such simple a priori truths as that if *x* is longer than *y,* then *y* is shorter than *x.* There will, then, be considerable overlap in the propositions ordinary rational persons rationally believe, particularly if they share the same environment and are similarly educated.

Implicit in the conception of theoretical rationality I am outlining is the idea that there should be a great deal of overlap in the rational beliefs of persons who experience the same phenomena or consider the same self-evident or even broadly a priori propositions.[27] I am assuming, of course, that we *can* experience the same colors and shapes, sounds and textures, tastes and smells, and the same kinds of pleasures and pains, and that we can consider the same a priori propositions, such as certain logical and mathematical ones. If this is so, then not only is there substantial overlap in the theoretically rational cognitions of sensorily and intellectually normal persons; we can also increase that overlap by the kind of positive communication constituted by testimony and decrease it by certain kinds of elaboration of our differences.

Closure Conditions for Rationality and Justification

There is a further question that arises when we consider the extent to which principles of rationality are properly modeled on those of logic. To philosophers, at least, it might seem that there should be at least this much parallel: just as logical entailment always preserves truth, logically valid inference always preserves rationality. If *p* is true and entails *q,* then *q* is true; if I rationally believe the former and validly infer the latter from it (and, as would be usual, hold it on the basis of the former), then I rationally believe the latter. I have already suggested that this *closure principle* (so called because it says that the class of rational beliefs is "closed" relative to the kind of inference specified) seems to hold for a great many cases. But it is not self-evident that there are no exceptions to it.[28]

The closure principle just formulated concerns *closure of rationality for inferential belief.* But our concern with the scope of theoretical rationality also extends to what propositions are *rational for a person* (to believe). We are interested not just in actual beliefs but also in theoretical rationality as applied to potential beliefs. In this light, we might hold that if one rationally believes that *p*, and *p* self-evidently entails *q*, then (other things equal) one *would be* rational in believing *q* on the basis of *p.* There are many closure principles that concern theoretical rationality. One is that if one has grounds on which believing *p* would be rational but one does not believe *p,* then, if *p* self-evidently entails *q,* one would be rational in believing *q* on the basis of *p* should one believe *p* on those grounds. This is plausible but not self-evident. What may be said with some confidence is that there are some appropriately qualified closure principles—including some that are inductive rather than deductive—that enable us to see a great number of propositions as theoretically rational *for* a person who has rational beliefs, or even just good grounds *for* rational beliefs, to start with.

Speaking more generally, we might say that for anyone with the range of theoretically rational elements that it is plausible to attribute to most people who have even a good grammar school education, theoretical rationality has indefinitely wide scope with respect to propositions that one can rationally believe by inferentially extending one's belief system. Every ground for a rational belief can render more than one belief based on it rational; every rational belief is a basis for inferences that can yield indefinitely many more rational beliefs.

end p.37

The Practical Authority of Theoretical Rationality

One further question should be briefly addressed here. How much scope does theoretical rationality have in practical matters? This question has aspects that we cannot approach here, but several points can be made briefly and will round out the treatment of theoretical rationality I am presenting. An extreme view—sometimes ascribed to Hume—is that there is no practical rationality, hence no particular ends we ought to seek in life; rather, action is guided by beliefs, and its success depends on whether it satisfies the agent's "basic" desires. Thus, if you want to fulfill your desires, you should try to

have rational beliefs to guide them, since these are more likely than nonrational ones to be true. One could reject this extreme view and hold instead that actions (and desires) are rational wholly on the basis of actual or potential rational beliefs. (The latter case may occur where one has the grounds for a belief that one should *A,* but has not formed the belief, which is at the time a *potential* rational belief). A more plausible position would be that an action is rational if and only if the person has grounds on which it is theoretically rational to believe that one may rationally perform it. This does not require that the action actually have a *basis* in theoretical reason. It also allows that a child can act rationally before having concepts adequate to form beliefs about rational action, as opposed to beliefs about means and ends. The point is only that practical rationality is a status that *can* be justifiably attributed to actions on the basis of theoretical reason.

A quite different view is that there are experiences, such as eating a delicious meal when hungry, that it is rational to want to have for their own sake, and there are actions connected with them, such as eating a delicious meal, that it is rational to perform for their own sake. Associated with this view is the position that it would not be rational to believe we should have such experiences if they were not already "worth wanting" and hence constitute appropriate grounds for the practical rationality of desire and action directed toward them. We need not assess all these ideas here. There are two points that would hold in any case regarding the scope of theoretical rationality in respect of its authority over practical reason. First, no one doubts that action and desire should be guided by theoretical reason, roughly in the sense that we should be guided in seeking our goals by *rational* means-ends beliefs. Second, few if any doubt that if we hold certain kinds of negative beliefs about an action, such as that performing it will be painful or will cause us to fail to get important things we reflectively want, then the would-be practical rationality of the action can be defeated.

The authority of theoretical reason over practical reason, then, is considerable. We cannot reach any destination without a route, and we cannot choose routes well unless we are guided by theoretically rational beliefs. On the other hand, we can have an excellent map without having a destination, and if none were worth visiting on its own account, why should we go anywhere? If nothing were worth

wanting or doing on account of what it is, why should we do anything? It seems unlikely that it would be rational to want to do it just on the basis of what we believe about it simply as a means to something else. This is a deep issue, however, particularly if we consider cases of actions required by morality. Fortunately, action and desire can receive support *both* from rational beliefs about them and from experiences of their intrinsically rewarding features or of sufficiently similar elements. Here we would have a case of wide scope for theoretical reason together with its cooperation with elements, such as enjoyable experiences, that support practical rationality in their own right.

At this point it is natural to ask whether a belief may be rational on a practical basis, as where one might be said to have a practical reason to hold it. One might, for instance, have excellent reason to think that believing one will survive a disease will help one do so. On some views, this is a pragmatic reason to believe that one will survive, and if we so regard it we might think such reasons can in some cases render a belief rational. [29] We

can distinguish, however, between a *reason for believing p* and a *reason for causing oneself to believe p.* It is true that causing oneself to believe *p* may *produce* a reason to believe it—as where one's believing one will survive the disease actually makes this likely—but once the basic distinction between the two kinds of reasons is observed, it seems doubtful that practical reasons of the kind in question—reasons for action—can double as theoretical reasons—reasons for belief. [30]

V. Theoretically Rational Persons

It might seem that once we understand theoretical rationality for individual cognitions, paradigmatically, beliefs, we can understand the notion of a (theoretically) rational person by simply specifying that a suitable proportion of the person's beliefs (or at least dispositions to form them) are rational and—depending on the person's experiences—perhaps also require beliefs of certain sorts.

Degrees of Rationality

Even if there is disagreement about the minimal proportion of rational beliefs required for (theoretical) rationality, we could at least define the notion of one person's being *more rational* than another (or than that person at a different time) in terms of the *number* of rational beliefs. But brief reflection shows that this will
end p.39

not do. For one thing, some beliefs are more important to one's rationality than others. One silly superstitious belief might be a mere stain on an otherwise reasonable cognitive record; a belief underlying the gambler's fallacy (which would have, say, a six on a fair toss of a die become more likely given its absence for a dozen successive tosses) can discolor large segments of that record. Moreover, even a large number of beliefs important in the relevant respects might exhibit little interconnection. Think of great mathematical knowledge isolated from beliefs permitting its applications (if this disconnection is even possible), or of a fine set of moral beliefs in the absence of related beliefs about human psychology. People with disconnected beliefs of these sorts might fail to be theoretically rational in an overall way.

Rational Integration

There is a kind of *cognitive integration* that is required in a person who is rational from the point of view of theoretical reason, as well as a twofold requirement: of an adequate proportion of rational cognitions and of the absence of certain kinds of "vitiating"

irrational beliefs, such as those that violate logical principles or prevent an appropriate response to experiential grounds of rational belief. There is no way to be quantitative here, but we may say that at one end is minimal theoretical rationality and at the other the kind that would be exhibited by a perfectly omniscient God.

Reasonableness

A theoretically rational person need not meet a high standard of rationality, say in exhibiting a critical mind or good judgment. In between minimal rationality and intellectual excellence is *reasonableness* in theoretical matters, a status above the former but not requiring satisfaction of the high standards essential for the latter. Similarly, a belief may be minimally rational, yet not reasonable, as where someone is influenced by arguments that, though not without plausibility, can be seen on careful reflection to be specious. For each of these cases, the relevant baseline depends heavily on the person's experience. The more limited a person's experience, the less in the way of rational belief we should expect the person to have, other things equal. But in a rational person there should still be an overall coherence not only within the system of beliefs, but also between it and the person's experience. When this pattern is combined with such intellectual assets as perceptiveness, good judgment, and a significant capacity for good reasoning, we may speak of a theoretically reasonable person.
end p.40

Global Rationality

Reasonableness in the theoretical domain does not entail global rationality, the kind that implies one's practical rationality as well. Even if certain beliefs imply motivation, there is no guarantee that a person reasonable in the domain of belief will have sufficient motivation—and appropriate emotions and attitudes—to qualify as a rational person overall. We cannot succeed as practical beings in the absence of theoretical rationality, but for practical success we need more. As to achieving theoretical rationality itself, true beliefs, no matter how numerous, are not sufficient; well-grounded beliefs, no matter how rich or insightful in content, do not imply it apart from integration, and even when integrated they may also fail to render a person rational overall. Logical powers, in the absence of suitably grounded beliefs to provide rationally held premises, are like an engine without fuel.
Understood in an overall sense, then, theoretical rationality requires the kind of well-groundedness of beliefs that is possible only given sensory and reflective experience as a basis; but an integration among the beliefs so grounded and the logical capacity to build inferentially beyond them are also needed. When theoretical rationality is well developed, the person will also have a measure of imagination, the kind that enables us to frame hypotheses, elaborate ideas, and even construct theories. But imagination, even if it normally bespeaks some degree of theoretical rationality, can also yield irrational beliefs

or hypotheses. The bad as well as the good can emerge from good grounds, but there is no limit to what can be built from them, nor any fixed direction in which rational speculation and imaginative flights may go. Theoretical rationality entails some degree of connection between our beliefs and basic sources, and it requires some integration among the elements that develop, at however great a distance, from them, but these constraints are not rigid. Theoretical rationality is compatible with many different kinds of content; it can burgeon in people with many different sorts of psychological dispositions; and it can improve indefinitely over time.

NOTES

An earlier version of this essay was given at the Universities of Frankfurt and Rome, and I am grateful for extensive audience discussions on both occasions. I also want to thank Alfred Mele for helpful comments on an earlier draft.

1. I have developed a detailed and comprehensive theory of rationality, applicable to practical as well as theoretical reason, in Audi 2001 , and some of what I say here is drawn from that book and defended there.

2. A longer treatment of theoretical rationality would also consider conditions un der which *change of belief* is rational, e.g., ceasing to believe a proposition in favor of a different one. For some philosophers, such change is the primary focus of rationality, roughly in the sense that belief change, but not "ongoing" belief, characteristically stands in need of justification. One question here is whether change of belief is a kind of action. If so, it should be governed by standards of practical reason; if not, then arguably it is rational when the person has better reason for holding a new belief than one it would replace. A detailed account of rational belief change that makes that notion epistemologically central is offered in Levi 1991 . For pertinent discussion of conditions for rational belief change, especially through making inferences, see Harman 1999 , especially chaps. 1 and 4. Also relevant are Kaplan 1996 and van Fraassen 1984 .

3. "Experience and reason" is a phrase often used by Roderick M. Chisholm among others; see, e.g., Chisholm 1966 , 59.

4. If "blind sight" is a case of perception, this may not be so (though it is arguable that the subject simply does not believe there are visual sensations or any other experiential element corresponding to perception).

5. An interesting question that arises here is whether *perceptual* consciousness, which has an external object, can be, except in a hybrid way, a mental state. For a case that it (and indeed knowing in general) can be a mental state, see Williamson 2000 . What is said about rationality in this essay is largely neutral with respect to that issue; but I take it that the view that something internal to the mind is what grounds rationality is consistent with the view that consciousness of external objects, whether a purely mental state or not, is direct.

6. The apparent noncausal character of abstract entities is a main reason that knowledge of them—indeed their very existence—is often considered problematic. For one kind of challenge to the causal inertness claim, see Plantinga 1993a , 115–17.

7. For introspection and consciousness, as for external perception, one can devise a plausible adverbial view, as described in Audi 1998 , chap. 1.
8. See Dretske 1981 and Alston 1991 for indications of how broad the notion of perception is.
9. The need for "if" here has been suggested already: a duplicate of me would, at the moment of creation, know dispositionally a great deal I now know from memory (not all of it, of course, because some depends on my actual history and it has no history as yet); but it is unclear how this depends on memory. Perhaps we should say that it does not depend on *remembering*—hence does not require the *operation* of memory—but does depend on *memorial capacity,* since it would not be true of me that if I needed to bring a certain item of knowledge to mind I would, unless I have sufficient memorial capacity to retain it from the moment of need, e.g., seeking a phone number I want, to the "next" moment, at which I bring it to mind.
10. Granted, I could memorially believe *p* but not know it (having too little evidence, say) and then be told by you that *p.* But if I now know it, this is on the basis of your testimony; I don't know it from memory until I retain the knowledge and not just the belief. Believing from memory can instantaneously become knowing, but does not instantaneously become knowledge from memory.
11. For a detailed discussion of the epistemology of memory, with many references to relevant literature, see Audi 1995b .
12. Consider, e.g., Hume's extraordinary affirmation of privileged access in the *Treatise*—"Since all actions and sensations of the mind are known to us by conscious

ness, they must necessarily appear in every particular what they are, and be what they appear" (1978, book 1, part 4, 190). This double-barreled claim is discussed in detail in Audi 1998 , chap. 3.
13. The relevant kind of understanding and the notions of a priori knowledge and justification in general are discussed in detail in Audi 1998 , chap. 4, and Audi 1999b .
14. For instance, one might look at a clock that one has reason to think is running. Suppose one knows it is *about* ten o'clock. If it is ten just as one looks at the clock, one might have a justified true belief that it is just ten, but does not know this. A brief treatment of such cases and many references to the literature are given in Audi 1998 , chap. 8.
15. There is a large literature on (Bertrand) Russell's paradox and on the theories of types devised, initially by him, to deal with it. A short account is provided in Barker 1964 , 83–89.
16. Thus, for God or any being with infinite memorial capacity, no use of reason *essentially* depends on the exercise of memory. I might add that even if the points made here about inference and memory are mistaken, the overall point that reason may ground justification for *p* without yielding knowledge of it can be illustrated by many other cases, presumably including that proposition that some classes are members of themselves (since this embodies a type-error).
17. This point must be qualified if W. V. Quine is right in denying that there is a viable distinction between the empirical and the a priori—at least one would have to speak in

terms of, say, differences in degree. For extensive criticism of Quine, see BonJour 1998 , and for the notion of a priori justification see also Audi 1999b .
18. For some major accounts see Harman 1973 , Lehrer 1974 , Davidson 1983 , and BonJour 1985 ; and for much critical discussion see Bender 1989 . It should be noted that in BonJour 1999 , BonJour has since abandoned coherentism.
19. This is suggested and to some degree argued in Audi 1998 and 2001.
20. That we do not form beliefs of all the kinds we are sometimes thought to form—particularly all those we would have if we believed whatever we would readily assent to upon simply considering it—is argued in detail in Audi 1994 .
21. If it is taken to be an internal relation among beliefs, their content does not matter, nor does their fit with experience. This sort of thing has been widely noted; see Moser 1993 and Bender 1989 for some relevant points and many references.
22. This is defended in Audi 1994 .
23. For a detailed critique of coherentist theories supportive of the points made here, see Plantinga 1993b and Bender 1989 .
24. For a statement of reliabilism, see Goldman 1986 ; for accounts of virtue epistemology see, e.g., Sosa 1991 , Zagzebski 1996 , and Greco 2000 . A brief statement of contextualism is given in DeRose 1992 . Audi 2001 makes it clear how each of these kinds of perspective is compatible with a moderate foundationalism.
25. The notion of degree of belief is treated in detail in Levi 1991 , Kaplan 1996 , Harman 1999 , and Joyce, chap. 10, this volume.
26. This issue is discussed in Audi 1997 . A contrasting view is developed in Fricker 2002 .
27. Self-evidence is analyzed and distinguished from other cases of the a priori in Audi 1999b .
end p.43

28. For my own attempt to show that there are exceptions, see Audi 1995a . For supporting works see Dretske 1970 , Nozick 1981 , and Klein 1995 .
29. For one kind of case for the possibility that practical reasons can support the rationality of belief, see Foley 1993.
30. This issue is explored, and the suggested conclusion defended, in Audi 1999a .
end p.44

Chapter 3
PRACTICAL ASPECTS OF THEORETICAL REASONING
Gilbert Harman

Albert thinks about what route to take to get to Boston. He thinks that, while the direct western route is faster, the scenic eastern route is longer but more enjoyable with less traffic. He is in a bit of a hurry but could probably arrive on time going either way. He eventually reaches a decision.

The reasoning Albert goes through in settling on what route to take is *practical.* He is deciding what to do.
At about the same time, Albert's friend Betty tries to decide what route Albert will take. She thinks about what Albert has done before, what Albert likes in a route, and how much of a hurry Albert is in. Betty's reasoning is *theoretical.* She is trying to arrive at a belief about what Albert will do.
Practical reasoning in this more or less technical sense leads to (or modifies) intentions, plans, and decisions. Theoretical reasoning in the corresponding technical sense leads to (or modifies) beliefs and expectations. There is also the possibility that reasoning of either sort leaves things unchanged.
Any given instance of reasoning may combine both theoretical and practical reasoning. In deciding which route to take, Albert may have to reach theoretical conclusions about how long it will take to go by the eastern route. In thinking about which route Albert will take, Betty may have to reason practically about whether to check her records about Albert's past trips to Boston.
end p.45

Nevertheless, there is a difference between theoretical reasoning and practical reasoning and a corresponding difference between theoretical reasons and practical reasons. In particular, there is a distinction between theoretical reasons to believe something and practical reasons to believe something. For example, Samantha has theoretical reasons to believe that knowledge of the history of philosophy is not very useful in actually doing good philosophy today, reasons based on a careful study of the history of philosophy and of the best recent philosophical literature. On the other hand, she has practical reasons to believe that knowledge of the history of philosophy is very useful in actually doing philosophy today, because she wants to be hired by a philosophy department that has a policy of only hiring candidates who believe that a solid knowledge of the history of philosophy is very useful to anyone who tries to do philosophy today.
A purely theoretical reason to believe something is sometimes called an *epistemic* reason to believe it, in contrast with a *nonepistemic* practical reason (Foley 1987).
There are interesting questions about how and to what extent practical reasons might be relevant to theoretical reasoning, strictly so-called. Practical reasons are certainly relevant to whether to undertake theoretical reasoning about a particular subject. Practical considerations may also be reflected in the role played by conservatism and simplicity in theoretical reasoning.

Preliminaries

Before discussing these issues, I need to discuss some preliminary points.

Reasoning and “Logic”

There is a use of the term “logic” to mean the theory of reasoning or inquiry, as in Hegel's *Logic* (1812) or Mill's *Logic* (1869). But in contemporary philosophy the term “logic” is often used for a theory of implication and inconsistency, as in accounts of truth-functional logic, quantificational logic, and modal logic. Terminology is not important, but it is important not to confuse issues about reasoning and inquiry with issues about implication and inconsistency.

Reasoning or inquiry is a process by which you change (or don't change) your views. A theory of reasoning or inquiry is a descriptive or normative theory of that process. The theory of implication and consistency concerns abstract prop erties of propositions and abstract relations between propositions. That is not an especially normative subject and it does not have an especially psychological subject matter. We can meaningfully ask whether the theory of implication and consistency has any special relevance to the theory of reasoning and inquiry, a question that is often hidden from view by the ambiguity of the term “logic.”

Reasoning or inquiry should not be identified with the construction of an argument or proof, although it may sometimes involve such construction. Even when reasoning does lead to the construction of an argument or proof, the process of reasoning does not normally begin by first considering the premises, then moving through intermediate steps, and finally ending with the conclusion. Anyone who has taken elementary geometry knows that proofs or arguments are sometimes constructed backward, from the conclusion through intermediate steps in reverse order to the premises. More often, you start in the middle and move both backward and forward in constructing an argument. Furthermore, when reasoning involves the construction of a proof or argument, the conclusion of your reasoning isn't typically the same as the conclusion of your argument. For example, in inference to the best explanation, your conclusion may be the “premise” of an explanatory argument whose “conclusion” is something that you started out believing and that the argument serves to explain.

An argument or proof is an abstract structure of propositions, consisting of initial premises, intermediate steps, and final conclusion. A formal system of proof might state certain misnamed “rules of inference” and require that each step in an argument should either be a premise or should follow from previous steps in accordance with one of the so-called rules of inference. Such “rules” are about implication, not inference, and they are “rules” only in the sense that they are constraints on what structures count as formal arguments in that system. They are rules that proofs must satisfy, not rules for reasoners to follow.

A valid proof or argument shows some of the implications of the premises. Of course, the premises imply themselves, so a typical proof shows additional implications. Reasoning, on the other hand, does not just add further conclusions to things you already accept. It typically also involves giving up some of things previously accepted. If we describe what you initially accept as “premises,” then we have to say that reasoning often involves abandoning premises and not just accepting further conclusions. (But it is best not to use the term “premise” in discussing reasoning, because of this term's association with arguments and proofs.)

It is not easy to specify a special connection between reasoning and the theory of implication and consistency. For example, although sometimes the fact that your prior

beliefs logically imply a conclusion may give you a reason to accept that conclusion, this does not hold in the general case. For one thing, you may not realize that the implication holds. For another, even if you do recognize the implication, the conclusion may be implausible, so that the implication may give
end p.47

you a reason to reject a prior belief rather than a reason to accept the conclusion (and it may not be true that there is a particular prior belief that one have a reason to reject). Even when the conclusion is not antecedently implausible, you will in the general case have no reason to be interested in whether it is true and so no reason to add it to your beliefs.

Distinguishing Theoretical from Practical Reasoning

I started this chapter with a contrast between Albert trying to decide which route to take (practical reasoning) and Betty trying to decide which route Albert is taking (theoretical reasoning). These examples suggest that to a first approximation theoretical reasoning is concerned with deciding what to believe and practical reasoning is concerned with deciding what to do. To a second approximation, we can say that theoretical reasoning is a process by which *in the first instance* you change your beliefs and expectations and that practical reasoning is a process by which *in the first instance* you change your choices, plans, and intentions. We have to say something like "in the first instance" because changing what you plan to do can affect what you believe will happen and changing your beliefs may lead you to change your plans.

There are obvious similarities between theoretical and practical reasoning. In both cases you start with antecedent beliefs and intentions and reason in a way that makes changes in those beliefs and intentions typically by subtracting some and adding others. (In the limiting case reasoning leaves things as they were at the beginning, with no change.)

But there are also important differences between theoretical and practical reasoning. A very important difference has to do with wishful thinking, which is perfectly proper in practical reasoning in a way that it is not proper in theoretical reasoning. Albert's preference for the eastern route can give him a practical reason to take the eastern route rather than the western route. But Betty's preference for Albert to be taking the eastern route does not in the same way give her a theoretical reason to believe that he is taking the eastern route.

Another important difference between theoretical and practical reasoning has to do with the reasonableness of arbitrary choices. Suppose Albert is trying to decide whether to take the eastern route or the western route and he finds that nothing favors one route over the other. Then it is reasonable for him to decide arbitrarily to take one of the two routes. If it is urgent for him to get to Boston, it would be a mistake for him to suspend judgment in this case. On the other hand, if Betty is trying to decide which route Albert is taking and there is no particular reason to think he is going one way rather than the other, it is *not* reasonable for her to decide arbitrarily that he is taking the one route rather than

end p.48

the other. In the theoretical case, Betty should suspend judgment. In the practical case, Albert's suspending decision can be deeply irrational.

The point about wishful thinking indicates a way in which a practical consideration deriving from your goals and desires is *not* properly relevant to your theoretical reasoning. But there are other ways in which practical considerations of this sort are properly relevant to your theoretical reasoning.

Practical Reasons to Reason Theoretically

Betty may have practical reasons to intercept Albert before Albert gets to Boston, so Betty may have practical reasons to figure out whether Albert is taking the eastern route or the western route. This illustrates one obvious way in which practical considerations can be relevant to theoretical reasoning—namely by being relevant to what to reason theoretically about.

A related point is that reasoning uses resources like time and concentration (Simon 1957 ; Gigerenzer et al. 1999). You have limited resources and reasoning about one issue keeps you from considering another. So, you have practical reasons to consider only certain questions rather than others. Practical considerations are also relevant to how much effort you should devote to investigating a given issue.

If Betty didn't have a reason to care which way Albert was going home, she would not have a reason to think about which way he was going home and she would not have a reason to reach any conclusion about which way he was going home.

For example, as already mentioned, the fact that Betty believes things that logically imply a given conclusion does not mean that she has sufficient reason to believe that conclusion. She may have no reason to be interested in whether that conclusion is true and every reason to be thinking about something else. Betty's beliefs logically imply infinitely many conclusions, most of absolutely no interest to her. She has to decide where to devote her resources. She should not clutter her mind with trivial consequences of her beliefs.

Suppose David points out to Betty that certain of her beliefs about roads in and near Princeton cannot all be true. Betty believes that Route 1 runs north-south, that Nassau Street runs east-west, and that Route 1 is parallel to Nassau Street. On discovering this conflict in her beliefs, Betty is not rationally required to drop everything to figure out which to abandon. She may have better things

end p.49

to do with her time. Maybe she should have lunch first. Maybe she simply has no reason to care that her beliefs are inconsistent in this way. [1]

Given resource limits, practical considerations are relevant to how much in the way of resources to devote to a given inquiry and to when to end an inquiry. The police have to

decide which cases to investigate, how much effort to put into each investigation, which cases to keep open and which to close. A scientific researcher faces the same question of where to devote resources. So do the rest of us all the time.
To reach a conclusion is, among other things, to conclude an investigation. Practical reasons are relevant to reaching a conclusion, at least to the extent that they are relevant to whether to stop devoting resources to that investigation. This is not to say that practical reasons can properly be used to decide between several competing theoretical conclusions. But practical reasons can properly be relevant to whether to end inquiry, for example on the grounds that further investigation is not likely to be worth the effort.

Conservation, Simplicity, and Coherence

Relevant factors in theoretical reasoning include conservatism, simplicity, and coherence. Roughly speaking, starting with an initial view, you try to retain as much as possible of that initial view (conservatism), to favor simpler over more complex hypotheses (simplicity), to reduce inconsistency (negative coherence), and to find explanations of things in which you are interested (positive coherence). [2]
Someone might ask what justifies a reliance on such factors as conservatism, simplicity, and coherence in our theoretical reasoning. Perhaps such reliance involves the sort of wishful thinking that we normally suppose is not theoretically reasonable. Maybe it's just that we want our present views to be correct and we don't want to have to change our minds. And maybe we want the general principles and theories we accept to be relatively simple because we have an aesthetic preference for simplicity or because it is easier for us to use simpler theories.
Perhaps reliance on conservatism, simplicity, and coherence can be justified as promoting our goals, in the way that believing in the usefulness of the history of philosophy might promote Samantha's goal of being hired by the Mooseton philosophy department. But then our reasoning would seem to be practical rather than theoretical, because the relevant considerations would be practical, not purely epistemic.
end p.50

I now want to look in more detail at simplicity and conservatism in order to assess this sort of worry.

Simplicity

The first point then is that theoretical reasoning often favors simpler hypotheses over more complex ones. For example, suppose we have reason to believe that some quantity *y* is a function of a quantity *x,* and we are trying to figure out what the function is, given data about particular cases. We are trying to discover the function *f* such that *y = f(x).*

If we have quite a bit of data and a linear function (of the form $y = f(x) = ax + b$ for constant a and b) fits the data, then even though there are also infinitely many more complex functions that also fit, we will be much more inclined to believe that the function is linear than that it is one of the more complex functions.

What explains this inclination toward simpler hypotheses? It is not exactly that we assume or presuppose that the world is simple. Our inclination is to accept the simpler of two hypotheses that account equally well for the data. Data can and will lead us to reject all of the absolutely simplest hypotheses. Our use of simplicity can and will sometimes lead us to accept very complex hypotheses, having rejected all the simpler ones.

Roughly speaking, we reason as if we accepted a conditional probability function p such that, if hypothesis h_1 is simpler than h_2 and e is our evidence, then $p(h_1|e) > p(h_2|e)$. I say that is only "roughly speaking" because in real life we trade off simplicity and data coverage. We allow for measurement error and noise in the data. Often, none of the hypotheses we consider fits the data perfectly. We weigh the extent to which a hypothesis fails to fit the data (perhaps as measured by the sum of squared error) against its complexity (measured in some way or other), trying to minimize some function of these two quantities. This means that our present evidential data e will often in practice be actually inconsistent with the hypothesis we end up inferring from that data, so strictly speaking the relevant conditional probabilities will be zero and the wording in the preceding paragraph is inaccurate.

Still, we favor a simpler hypothesis over infinitely many more complicated hypotheses that do equally well or better at data coverage. We reason as if we believed that the simpler hypothesis is more likely to be correct in this case. But why should we believe this?

Actually, we do not exactly "believe" this. Our preference for simplicity is "built into" our system of reasoning—as it were, part of our initial probability distribution. It is a basic aspect of our epistemic probability. [3] Our inferential practice treats simpler hypotheses as more epistemically probable than corresponding more complex hypotheses that account equally well for the data.

Inductive Bias

Once we realize that we are influenced by simplicity in this way, we can ask whether we should continue to allow ourselves to be influenced. We can ask why we should go along with this tendency in our reasoning practices.

One thing that seems relevant is that a reasoning system needs inductive bias if it is to reach any inductive conclusions at all. [4] A system without inductive bias cannot learn from experience. Now, perhaps certain entities can survive without learning, but ordinary people cannot. Some bias that will enable learning has to be built into us somehow, perhaps as the result of evolution by natural selection.

Example

Go back to the example in which we have reason to believe that y is a function of x and would like to know what the function is. We might proceed as follows.
Consider the set F of functions that can be expressed in standard mathematical notation. There are countably many such functions, which means they can be ordered in a way that is correlated with the natural numbers. Choose one such ordering, $f_1, f_2, \ldots, f_n, \ldots$. Using that ordering, we get some data and select the earliest function in the ordering that is compatible with the data. We then get more data and see whether the chosen function fits the additional data. If not, we again choose the earliest function on the list that is compatible with all the data we now have. If the right function is included in our list, F, this method will eventually arrive at that function, given enough time (Putnam 1963).
There is the complication that we must allow for noise in the data. So, we must make some sort of trade-off, for example, choosing that function f_n for which the sum of n and the mean squared-error on the data is least.
Our initial ordering of hypotheses might be based on how easy it is to use a hypothesis to answer questions in which we are interested. This would be one sort of simplicity ordering, where "simple" means simple to use. [5]

Ordering *Sets* of Hypotheses

Alternatively, we might rank all linear functions ahead of all quadratic functions, for example. In that case, we would not have a simple well-ordering of hypotheses,
end p.52

since infinitely many hypotheses would come between any linear hypothesis and any nonlinear quadratic hypothesis. Instead, we might use an ordering of cumulative *sets* of functions, $F_1, F_2, \ldots, F_n, \ldots$ where $F_1 \subset F_2 \subset F_3 \ldots$.
The idea is then first to consider the hypothesis in each of these sets that has the least squared error on the data. For each such hypothesis, we trade off the amount of the error for that hypothesis against the complexity of the earliest of the hypothesis classes to which it belongs.
We would presumably order these sets of functions so that later sets can accommodate more data points with no more error than earlier sets do. For example, given any two data points there is some linear hypothesis that fits them perfectly, so it is not interesting that some linear hypothesis fits two data points. However, it is not true that for any three data points there is a linear hypothesis that fits them perfectly. So, it is somewhat interesting if we have only three data points and some linear hypothesis fits them exactly or fairly closely. On the other hand, it is not interesting that some quadratic function captures three data points exactly, but perhaps mildly interesting that some quadratic function captures four data points.
Suppose more precisely that we choose classes of hypotheses $F_1 \subset F_2 \subset F_3 \ldots$ in such a way that (1) for any N data points there is some hypothesis in F_{N+1} that exactly fits the points, but (2) it is not generally the case that for any N data points there is some hypothesis in F_N that fits the points. Again we would have a simplicity ordering of sorts.

Using this measure of simplicity when we trade off simplicity against error might (at least under certain conditions) promise eventually to lead us to accept a hypothesis with a relatively low average error. [6]

Practical Reasons to Be Sensitive to Simplicity

Our inferential practices show a bias toward simpler hypotheses. We are considering the question whether we ought to continue to allow this bias.

Perhaps we can say that simpler theories are more likely to be true than more complicated theories that account for the same data. We can say this, anyway, if likelihood is epistemic likelihood, because considerations of simplicity are built into the procedures that determine what is likely in that sense.

On the other hand, there seems to be no direct noncircular argument that simpler theories are more likely to be true in the sense of objective likelihood. (But we will come back to this issue.)

We might be able to argue for an inductive bias that rests on one or the other of the two types of simplicity we have considered. The argument might be that that sort of inductive bias promises to help us eventually find the answer to

end p.53

questions we are interested in answering or that it promises to help us eventually to choose a hypothesis with as low an error rate as possible while having other practical advantages.

These are practical reasons to acquiesce in an inductive bias that favors simplicity, not theoretical reasons. But they do not reduce theoretical reasoning to practical reasoning nor do they build wishful thinking into theoretical reasoning—at least not in the sense that *within* theoretical reasoning the desirability of a conclusion counts as a reason to believe it. There can be practical reasons for designing or acquiescing in a system that does not allow wishful thinking. A system defended in this way need not allow the internal use of practical considerations to decide between competing hypotheses.

Conservatism

Let us briefly look at an analogous issue concerning the role of conservatism in theoretical reasoning. Our reasoning is conservative in the sense that we start with our present view and try to improve it by getting rid of inconsistency and by increasing its coherence in ways that help us answer questions in which we are interested (Rawls 1971 ; Goodman 1965 ; Peirce 1931–58 ; Quine and Ullian 1970 ; Dewey 1938).

So, there is a further bias in reasoning beyond a simple inductive bias. This further bias favors beliefs that we already have over propositions that we do not already accept.

Again we can ask whether we should acquiesce in this further conservative bias. An alternative idea would restrict the conservative bias to certain “foundational” beliefs, such as beliefs about your most immediate experience and beliefs based on the recognition of self-evident truths (Descartes 1637 ; Foley 1987; Alston 1989 ; Chisholm 1982).
But there is a compelling argument for conservatism and against special foundationalism—namely that special foundationalism leads inevitably to scepticism, and that again one will not be able to learn much of anything if one cannot rely on one's other nonfoundational beliefs (Harman 2003). Here is another practical reason to acquiesce in a certain way of doing theoretical reasoning.
Again, this sort of practical reason for an aspect of theoretical reasoning does not imply that theoretical reasoning itself is practical reasoning and it does not imply that theoretical reasoning involves wishful thinking because of its bias toward conservatism.
end p.54

Nonpractical Interpretation

We can also interpret the reasons offered for simplicity and conservatism in a way that does not treat them as practical reasons at all. Instead, we can think of these considerations as showing that a system without inductive bias in favor of simpler hypotheses or without a bias toward conservatism would not be able to reach conclusions that ought to be reached.
The point is that it would be deeply irrational to reason in a way that leads to skepticism. Any skeptical system simply gives the wrong results about what one should believe.
Someone might object that this line of thought is circular. I respond that it is not circular—it is reasoning! To think there is circularity here is to think of reasoning as the production of a logical argument with premises, intermediate steps, and a conclusion. For then it seems that the relevant premises must include the rationality of a system with inductive and conservative bias. But reasoning is not to be identified with the production of an argument. Reasoning is reasoned change in view in a way that improves coherence in a way that helps to answer questions in which you are interested.
Conclusion
My brief conclusion is that, although wishful thinking is not relevant in theoretical reasoning in the way that it is relevant in practical reasoning, certain aspects of theoretical reasoning can be given a practical defense. That defense does not mean that wishful thinking is allowed internally to theoretical reasoning. The defense can also be given a nonpractical interpretation in terms of what conclusions ought to be reached.

NOTES

1. David Lewis noticed that many people who live in Princeton believe these three things. When this is pointed out to them, most Princeton residents are amused but not motivated

to correct their beliefs. (It turns out that Route 1 actually runs northeast-southwest when it is near Princeton, not north-south, and is anyway not really parallel
end p.55

to Nassau Street, which does run east-west. I have tried to explain this to my neighbors but they do not care.)
2. Pollock 1979 distinguishes positive and negative coherence in this way. Simplicity might be included in positive coherence and not counted as a distinct factor.
3. *Epistemic probability,* which has to do with what it is more or less reasonable for one to believe via theoretical reasoning, is to be distinguished from *objective probability,* which has to do with frequencies or propensities that one may or may not be aware of. Given a normal looking six-sided die, one might reasonably suppose that any of the sides is equally likely to be topmost after the next roll of the die. That equal likelihood is a matter of epistemic probability. In fact, the die may be constructed in such a way that there is a strong propensity for the side labeled "4" to come up. The objective probability of 4 coming up may be 0.8 even though the epistemic probability of getting a 4 on the next roll is 1/6 or 0.17.
4. This obvious point is discussed in the context of machine learning in Mitchell 1997 , chaps. 1–2. It is one moral of Goodman's (1965) "new riddle of induction," as discussed by philosophers. Stalker 1994 is a collection of essays on that riddle.
5. Ludlow 1998 defends this view of simplicity, citing earlier versions in Peirce 1931–58 and Mach 1960 . Harman 1994 argues for a similar view after surveying alternatives. Related computational approaches to simplicity are defended in Angluin and Smith 1983 , Blum and Blum 1975 , Blum 1967 , Gold 1967 , Kugel 1977 , Solomonoff 1964 , Turney 1988 , and Valient 1979 . Sober 1975 argues for a "semantic" interpretation of simplicity; later, Sober 1988 and Sober 1990 argue against the general relevance to inference of certain notions of simplicity.
6. For further discussion, see e.g., Mitchell 1997 , chap. 7.

Chapter 4
PROCEDURAL AND SUBSTANTIVE PRACTICAL RATIONALITY

Brad Hooker
Bart Streumer

According to many philosophers, all practical rationality is *procedural.* According to other philosophers, besides procedural practical rationality, there is also a different kind of practical rationality, which is *substantive.* This chapter is about the debate between these two groups of philosophers, whom we shall call *proceduralists* and *substantivists* (see also Smith, chap. 5, O'Neill, chap. 6, and McNaughton and Rawling, chap. 7, this volume).
In section 1, we explain the distinction between procedural and substantive practical rationality. In section 2, we outline the view of David Hume, who is often seen as the first

proceduralist. In section 3, we outline Richard Brandt's modern defense of proceduralism. In section 4, we set out Bernard Williams's very influential arguments for proceduralism. In section 5, we discuss the main argument for substantivism. In section 6, we outline how substantivists could criticize Brandt's defense of proceduralism. In section 7, we set out how substantivists could criticize Williams's arguments for proceduralism. In section 8, we
end p.57

discuss the possibility of being a proceduralist about practical rationality, but a substantivist about practical reasons.

1. The Distinction between Proceduralism and Substantivism

Suppose that Jack has a disease from which he will die in thirty years' time, unless he takes a certain medicine now. If he takes this medicine, it will cure him completely, without any side effects. Jack knows all this, but he lacks the desire to take this medicine. According to one group of philosophers, Jack can be open to rational criticism for lacking this desire *only* if

(1) He has beliefs and other desires from which he can rationally reach the desire to take this medicine, but he fails to reach this desire. [1]

For example, suppose that Jack has the desire to get married next year, and has the belief that he cannot get married next year unless he takes this medicine. In that case, he can rationally reach the desire to take this medicine from the beliefs and desires that he has. [2] And in that case, according to these philosophers, Jack can be criticized for failing to be *procedurally* practically rational.
According to another group of philosophers, Jack can be open to rational criticism for lacking this desire if

(2) Whether or not he has beliefs and other desires from which he could rationally reach a desire to take this medicine, he fails to have this desire.

For example, suppose that Jack does not have a desire to get married next year, and does not have any other beliefs and desires from which he could rationally reach the desire to take this medicine. In that case, Jack cannot be criticized for failing to be procedurally practically rational. But, according to these philosophers, Jack *can* be criticized for failing to be *substantively* practically rational.
The first group of philosophers defend:

Proceduralism: An agent can be open to rational criticism for lacking a desire only if the agent can rationally reach this desire from the beliefs and desires that he or she has. [3]
end p.58

We shall call such philosophers *proceduralists.*
The second group of philosophers defend:
Substantivism: An agent can be open to rational criticism for lacking a desire whether or not the agent can rationally reach this desire from the beliefs and desires that he or she has. [4]
We shall call such philosophers *substantivists.*
Proceduralists usually make a distinction between instrumental and noninstrumental desires. Noninstrumental desires are our foundational desires, and all our other rational desires are instrumental to the fulfillment of these foundational desires. For example, if Jack does acquire the desire to take this medicine, this desire will probably be an instrumental desire. This desire could, for example, be instrumental to the fulfillment of a foundational desire to be healthy, or a foundational desire to lead a happy life.
Proceduralists and substantivists often formulate their views in terms of reasons rather than in terms of rational criticizability. That is, they formulate their views as:
Proceduralism: An agent can have a reason to have a desire only if the agent can rationally reach this desire from the beliefs and desires that he or she has.
Substantivism: An agent can have a reason to have a desire whether or not the agent can rationally reach this desire from the beliefs and desires that he or she has.
Many proceduralists and substantivists treat the formulations of proceduralism and substantivism in terms of reasons as equivalent to the formulations in terms of rational criticizability. Until the final section of this paper, we will also treat them as equivalent.

2. Hume's Proceduralism

Proceduralists often invoke David Hume as the first defender of their view. Hume famously wrote:
'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger.'Tis as little contrary to reason to prefer my
end p.59

own acknowledged lesser good to my greater, and have a more ardent affection for the former than the latter. (Hume [1739] 1978 , 416)
According to many proceduralists, what Hume is suggesting here is that practical rationality cannot require that we have certain desires when we cannot reach these desires from our present desires. Instead, on Hume's view, practical rationality is merely a matter of our desiring efficient means to the fulfillment of our noninstrumental desires, which are not themselves subject to rational appraisal. In this sense, many proceduralists claim, Hume's view is that all practical rationality is procedural. According to other

philosophers, however, Hume's view is that there is no such thing as practical rationality at all (see Darwall 1983 , 53; Hampton 1995 ; and Millgram 1995). Nevertheless, because it is often thought that, on Hume's view, all practical rationality is procedural, proceduralism is often called "Humeanism." [5]

Those who claim that Hume was a proceduralist usually think that, on Hume's view, there are two ways in which we can rationally reach a new desire from our present desires. The first way is:

(1) Acquiring a new desire for something that is a means to something else that we currently desire.

For example, suppose that Sarah desires to pass a certain examination. In order to pass this examination, she must study for it. Even if she does not yet have the desire to study for the examination, she can reach this desire from her present desire to pass the examination, because studying for the examination is a means to passing it. According to many proceduralists, on Hume's view, practical rationality can then require that Sarah has the desire to study for the examination, given that she has the desire to pass it. Because, in such cases, acquiring the new desire is *instrumental* to fulfilling a present desire, proceduralism is also often called "instrumentalism" (see, e.g., Fehige 2001).

The second way is:

(2) Rationally acquiring a new empirical belief that leads to acquiring a new desire. [6]

For example, suppose that John believes that most lawyers are poor. In fact, however, given the evidence that he already has, he should rationally believe that most lawyers are rich. Theoretical rationality therefore requires him to have the belief that most lawyers are rich. Suppose that, if John had the belief that most lawyers are rich, he would desire to study law. According to (2), he could then rationally reach the desire to study law from his present desires. And according to many proceduralists, on Hume's view, practical rationality can then require that he has the desire to study law.

Obviously, (1) and (2) are related. For we might ask, Why is it that, if John

knew that most lawyers are rich, he would desire to study law? Presumably, the answer is that John already has a desire to be rich. So the desire that John can rationally reach from his present desires according to (2) is also a desire that he can rationally reach from his present desires according to (1). Therefore, if the new beliefs that he acquires under (2) are all beliefs about means to achieve things that he already desires, (1) and (2) are equivalent.

3. Richard Brandt's Proceduralism

A prominent modern defender of proceduralism is Richard Brandt. On his theory of rationality, proceduralism is the view that an agent has a reason to perform a certain act if and only if this act will fulfill whatever desires the agent would have after he or she would have undergone what Brandt calls *cognitive psychotherapy* (Brandt 1979 , 11, 111–13). Cognitive psychotherapy on an agent consists of:

(a)Putting aside any of the agent's desires that are founded on nonempirical beliefs (such as normative beliefs).
(b)Subjecting the agent's remaining desires to full empirical information, which may expunge some of the agent's desires and elicit some new ones.
(c)Making sure the agent's reasoning is logically correct.

On Brandt's view, there cannot be anything rationally wrong with an agent's desires as long as they don't result either from logically invalid reasoning or from less than full empirical information (Brandt 1979 , 11, 111–13).

By way of illustration, suppose that Fred has a strong desire not to eat apples. First, suppose he does not want to eat apples because he believes that eating apples expresses rebellion against God. In this case, his desire not to eat apples is founded on a nonempirical belief (that one shouldn't rebel against God). Any desire founded on a nonempirical belief would be purged by cognitive psychotherapy. Brandt wants normative theory to start from desires that are not a product of normative beliefs (Brandt 1979 , 2, 3, 13; Brandt 1989 , 127). He wants rationality to *generate* normative beliefs, not to *presuppose* them.

Suppose that Fred's desire not to eat apples is not founded on any normative belief, but comes instead from an empirical belief. Suppose it comes from the empirical belief—which, as it happens, is false—that eating apples is likely to make him ill. Now, once Fred learns that eating apples is more likely to help keep him from getting ill, he might go from desiring not to eat apples to desiring to eat them. For Brandt, what it is rational for an agent to do depends on what the agent would desire if his or her empirical beliefs were correct. It does not depend on what the agent does actually desire when his or her desire is based on a false empirical belief.

Brandt does not maintain that all other-regarding desires are to be ignored. Contrast what might be called *natural* concern for others with what might be called *conscientious* concern for others. An agent has natural concern for others if his or her desire that others do well expresses an underived altruistic impulse in the agent's nature, as opposed to being derived from a desire to comply with moral duty, or derived from a desire for other ends. In contrast, an agent has conscientious concern for others to the extent that his or her desire that others do well comes from his or her normative belief that the agent is morally required to desire this, and is thus founded on a normative belief. Since Brandt's

theory of rationality puts aside desires founded on normative beliefs, it puts aside conscientious concern for others, but it does not put aside natural concern for others. Different people have different degrees of natural concern for others. According to Brandt, this is part of why it may be rational for one agent to do something that it would not be rational for another agent to do. If Laura has greater natural benevolent concern than Emily does, then it can be rational for Laura to make greater sacrifices for the benefit of others than it would be rational for Emily to make.

According to Brandt, one of the elements of cognitive psychotherapy is making sure that one's reasoning is logically correct. For example, suppose that George desires not to be in the presence of transsexuals. And suppose this desire developed in him as a consequence of his once meeting a transsexual. On that occasion, the person made a pass at him. George then made the hasty generalization that most or all transsexuals, when in his presence, would make a pass at him. Since he generalized on the basis of just one instance, George made a mistake in inductive reasoning. That was a mistake in reasoning, whether or not the belief he arrived at was true.

Suppose the truth is that transsexuals are no more likely than nontranssexuals to make a pass at George. So the belief George arrived at via hasty generalization was in fact false. Given that the belief he arrived at was false, then, contrary to what Brandt claims, whether or not George arrived at this belief via faulty reasoning is ultimately irrelevant. What matters is only whether his desire not to be in the presence of transsexuals would extinguish once he became fully aware that transsexuals are no more likely than nontranssexuals to make a pass at him.

Now suppose (what is only just conceivable) that the belief George arrived at via hasty generalization was in fact true—that is, that most or all transsexuals would (when in his presence) make a pass at him. Then, contrary to what Brandt suggests, that George arrived at this belief via a hasty generalization is, again, ultimately irrelevant. Rather, on Brandt's view, what matters is only what George

end p.62

would want after he had been vividly and repeatedly exposed to the relevant empirical facts.

Because of this point about the irrelevance for Brandt of the logical or illogical reasoning involved in the acquisition of desires, as well as the irrelevance of instrumental desires, Brandt's theory boils down to the following:

everyone has most reason to do whatever best fulfills the set of noninstrumental desires that he or she would have after maximum exposure to all relevant empirical facts, where this set of desires does not include any desires founded on normative beliefs.

This theory is to some extent *idealized* in that it grounds reasons for action not in the desires the agent happens to have now, but in the desires the agent would have after maximum exposure to empirical information. The theory is *empiricist* in that it eschews reference to normative facts or properties. The theory identifies good reasons for action as whatever would fulfill the agent's desires, not including desires resulting from normative beliefs.

4. Bernard Williams's Proceduralism

The most influential recent defender of proceduralism is Bernard Williams (Williams 1981 , 1995a , 1995b). Williams's arguments for proceduralism can be set out as follows (Hooker 1987).

Williams defines what he calls an agent's *subjective motivational set* as a set that includes the agent's present desires, plus the agent's "dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may abstractly be called, embodying commitments" (Williams 1981 , 105). He then defines *rational practical deliberation* as:

(a) ascertaining what way of satisfying some element in one's subjective motivational set would be best in the light of the other elements in the set, or
(b) deciding which among conflicting elements in one's subjective motivational set one attaches most weight to, or
(c) "finding constitutive solutions, such as deciding what would make for an entertaining evening, granted that one wants entertainment." (Williams 1981 , 104)

end p.63

Williams defines what he calls an agent's *internal* practical reasons as reasons that can come to motivate this agent if the agent engages in rational deliberation that starts from his or her subjective motivational set. And he defines *external* practical reasons as reasons of which it does not need to be true that they can come to motivate this agent if the agent engages in rational deliberation that starts from his or her subjective motivational set.

Williams then claims that there are no external practical reasons. In other words, he claims that practical rationality is procedural, in the sense given by his definition of subjective motivational set and (a), (b) and (c) above. [7] Williams has two main arguments for this view.

His first argument appeals to the role that claims about reasons play in the *explanation* of what people do. He writes: "If there are reasons for action, it must be that people sometimes act for those reasons, and if they do, their reasons must figure in some correct explanation of their action" (Williams 1981 , 102). Starting from this observation, Williams gives an argument that can be set out as follows:

(P1) It must be possible for a reason for doing something to explain why an agent does this thing.
(P2) A reason can explain why an agent does something only if this agent is motivated by this reason to do this thing.
(P3) An agent can be motivated by this reason to do this thing only if the agent either already believes that he or she has this reason or can come to believe that he or she

(P1)It must be possible for a reason for doing something to explain why an agent does this thing.
has this reason by rational deliberation. [8]
(P4)All reasons that an agent either already believes he or she has or can come to believe he or she has by rational deliberation are internal reasons.

So,

(C) All reasons are internal reasons.

Williams's second argument concerns the *content* of claims about reasons. He writes: "*What* is it that one comes to believe when he comes to believe that there is a reason for him to φ, if it is not the proposition, or something that entails the proposition, that if he deliberated rationally, he would be motivated to act appropriately?" (Williams 1981 , 109). [9]
Williams's argument here can be set out as follows:

(P1*)The only intelligible content of the claim that there is reason for an agent to φ is, or entails, that the agent would be motivated to φ if he or she deliberated rationally.
(P2*)The content of the claim that there is an external reason for an agent

end p.64

to φ cannot be, and cannot entail that, the agent would be motivated to φ if he or she deliberated rationally.
So,
(C*)The claim that there is an external reason for an agent to φ has no intelligible content.

Williams considers two possible replies that substantivists—or, as he calls them, "external reasons theorists"—might give to this argument.
The first reply that substantivists might give is that the claim that there is an external reason for an agent to φ means that this agent would be nicer, more considerate, more courageous and the like if he or she were to φ. Williams writes:
There are many things I can say to or about [a man who does not φ]: that he is ungrateful, inconsiderate, hard, sexist, nasty, selfish, brutal, and many other disadvantageous things.There is one specific thing that the external reasons theorist wants me to say, that the man has a reason to be nicer.But if that is thought to be appropriate, what is supposed to make it appropriate, as opposed to (or in addition to) all the other things that may be said? (Williams 1995a , 39)

Consider statements like “φ-ing is insensitive to the feelings of others,” “φ-ing is dishonest,” and “φ-ing is pernicious to society.” Such statements cannot be rejected merely on the grounds that an agent lacks any desire that will be served by this agent's avoiding insensitivity, dishonesty, or perniciousness. In this sense, the concepts “insensitive to the feelings of others,” “dishonest,” and “pernicious to society” are externalist ones. What Williams seems to be suggesting is that, since we already have these externalist concepts to deploy against culprits, it is hard to see how deploying externalist claims about *reasons* against culprits could add anything distinctive to statements like “φ-ing is insensitive to the feelings of others,” “φ-ing is dishonest” and “φ-ing is pernicious to society.”

At this point, it may be helpful to contrast Williams's position with Gilbert Harman's. [10] Consider the spectrum of concepts in figure 4.1 . Start with the box on the right end of the spectrum. Harman and Williams agree that whether specific evaluative concepts can be ascribed to an agent's action typically does not depend on the agent's desires. Suppose we are inclined to evaluate some agent's act as dishonest and harmful to others. We would not withdraw those evaluations when we learned that this agent doesn't disapprove of dishonesty or harming others and doesn't have a desire to avoid dishonesty or harming others.

Now consider the box on the left end of the spectrum in figure 4.1. Again, Williams and Harman agree. They both believe that whether an agent has good reason to do some act *does* depend upon the agent's desires. [11]

What Williams and Harman disagree about is the status of the middle category—that is, the status of moral verdicts. [12] Williams takes moral wrongness not

end p.65

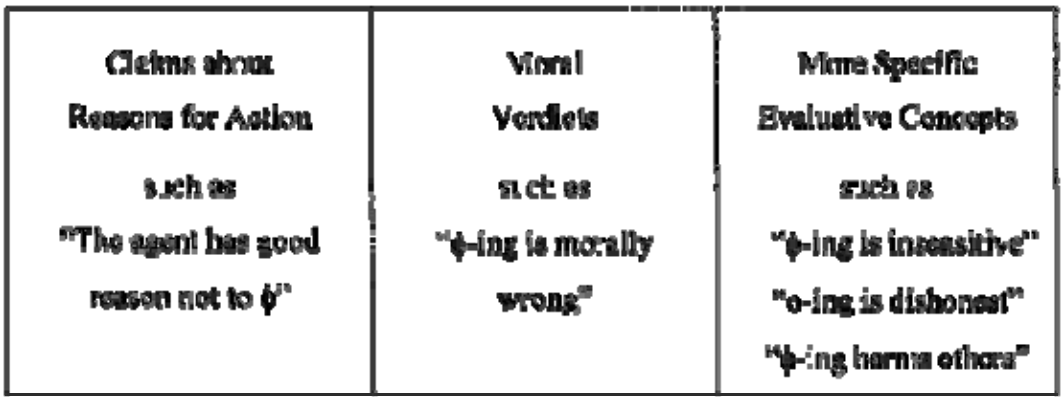

Figure 4.1

to be grounded in, and so not to be hostage to, the agent's desires. In contrast, Harman takes moral wrongness to be at least partly grounded in, and so to be hostage to, the agent's desires. [13] Harman thinks that an agent cannot have a reason not to φ unless the agent has a desire that not φ-ing would fulfill. He also thinks that φ-ing cannot be morally wrong unless the agent has a reason not to φ. Therefore, he thinks that φ-ing cannot be morally wrong unless the agent has a desire that not φ-ing would fulfill.

In short, neither Harman nor Williams relativize application of the *more specific evaluative concepts* to the agent's desires. Both Harman and Williams do relativize claims about *good reasons for action* to the agent's desires. Williams and Harman part company over whether *moral verdicts* should be relativized to the agent's desires.

Williams also considers a second reply that substantivists might give to his argument. According to this reply, the claim that there is an external reason for an agent to φ means that if this agent were a well-informed and well-disposed deliberator, he or she would be motivated in these circumstances to φ (Williams 1995b , 109).

Against this reply, Williams stresses that an agent may lack the dispositions and capacities that a well-informed and well-disposed deliberator would have. In such cases, Williams claims, it is implausible to say that what this agent has reason to do depends on the dispositions and capacities of a well-informed and well-disposed deliberator, rather than on the dispositions and capacities of this agent himself or herself. For example, suppose that Jane can't stop herself from drinking alcohol once she starts. In that case, it seems that Jane has a reason not to accept even one glass of alcohol, even though a well-informed and well-disposed deliberator (who would of course not have the disposition to drink too much once started) might have no such reason. Or suppose that, after a day of hard work, Tom is tired and irritable, to the point that Tom would probably end up picking a fight if he went to the pub. In that case, Tom has a reason not to go to the pub that a well-informed and well-disposed deliberator might not have, since a well-informed and well-disposed deliberator would not have the disposition to become irritable and pick fights when tired.

Substantivists could reply to this that the test for whether an agent has a reason to do something is whether a well-informed and well-disposed deliberator would be motivated *in these circumstances* to do this thing. If Jane can't stop herself once she tastes alcohol, and if Tom is irritable after a day of hard work, these facts are arguably part of the circumstances that Jane and Tom find themselves in. Therefore, *in the circumstances that Jane and Tom find themselves,* a well-informed and well-disposed deliberator *would* be motivated not to accept even one glass of alcohol and not to go to the pub.

Williams's rejoinder to this is that if we let such things count as differences in circumstances, we are acknowledging that differences in our affective states can make a difference to the reasons we have. Given his definition of an agent's subjective motivational set as including “dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projects, as they may abstractly be called, embodying commitments,” Williams suggests this puts us back on the road to the view that all reasons are internal.

These arguments lead Williams to suppose that claims about external reasons are mere empty rhetoric, or, as he puts it, “bluff” (Williams 1981 , 111). On Williams's view, internal reasons are the only reasons there are, and all practical rationality is procedural in the sense he has defined.

5. Arguments for Substantivism

The main form of argument for substantivism comes in the form of examples. Elizabeth Anscombe proposed this example. Suppose that someone who knows all relevant empirical facts and is reasoning logically nevertheless says that he wants a saucer of mud (Anscombe 1957 , 70). And suppose that this person does not want the mud as an artistic material for the creation of something else. Nor does he want it for throwing in someone's face. Nor does he want it as a symbol for something else. He claims simply to want the saucer of mud for itself. According to substantivists, such a noninstrumental desire would be rationally criticizable (or indeed unintelligible)—whether or not the person would abandon his desire for a saucer of mud if he deliberated in a procedurally rational way with full empirical information.

There are lots of other examples. Warren Quinn offered the example of someone with a disposition to turn on radios whenever possible (Quinn 1993 , 236). This person does not turn on radios because she wants to hear music or news or

other radio programs. Nor does the person want to turn them on in order to test them, or in order to disturb other people. The desire in question here is supposed to be a noninstrumental one. But, substantivists would say, such a desire seems rationally criticizable.

Such examples can be multiplied infinitely. For there is an infinite number of things that are not worthy of being desired as ends *in their own right* (however instrumentally desirable they may be in some situations as means to or as symbols of other things). When these things are not means to or symbols of other things, there is no "desirability characterisation" of them (Anscombe 1957 , 72–73). According to substantivists, when something has *no* feature that would make the thing worth desiring as an end in itself, then desiring this thing for its own sake is rationally criticizable whether or not a desire for this thing would result from logically valid deliberation from the agent's desires.

Other examples of desires that substantivists would claim are rationally criticizable involve irrational patterns. Consider Derek Parfit's example of "Future Tuesday Indifference" (Parfit 1984 , 124). Suppose that someone cares equally about what happens to her on every day of the week except Tuesdays and that she cares not at all what happens to her on any future Tuesday. She does care on Tuesday about what is happening to her on that day, but she never cares about what happens to her on future Tuesdays. In addition, suppose that this person does not take Tuesdays to be symbolic of something important, and does not have any other special beliefs about Tuesdays. She is indifferent to her future Tuesdays for no further reason than that they are Tuesdays. This person is clearly irrational, substantivists would say. The fact that a harm or benefit occurs on a Tuesday is no reason to discount it. The fact that some days are Tuesdays is not really a reason to accord them lesser (or greater) significance. Indifference to future Tuesdays involves drawing a line arbitrarily and unnecessarily. Again, there are an infinite number of examples of the same general kind. These are examples in which a pattern of concern discounts what happens during some unit of time, or space, for no further reason.

Being indifferent to what happens on future Fridays, for no further reason than that they are Fridays, would be just as irrational as indifference to what happens on future

Tuesdays, according to substantivists. Likewise, to be indifferent to what happens on any future day between 4:02 and 4:05, for no further reason than that those times are those times, would also be irrational according to substantivists. Or imagine someone who cares deeply about what happens to those less than one mile away from her but not at all about anyone who is more than one mile away (Parfit 1984 , 125). Again, suppose that the distinction between those within a mile and those outside a mile is made for no further reason. (It is not that all members of this person's family are within a mile.) According to substantivists, such "within-a-mile altruism" makes an unnecessary and arbitrary dis
end p.68

tinction and is therefore rationally criticizable, whether or not this person could abandon this pattern of concern on further reflection.
Another example is the far more familiar one that anyone with absolutely no concern for his or her own future well-being would be irrational, according to substantivists. [14] Consider the fifteen-year-old who says, "I don't care one bit about anything that happens to me after I'm thirty." According to substantivists, this person fails to care about something she should care about, and is to that extent rationally criticizable, whether or not she would come to care about her further future if she thought clearly about it with full empirical information (Sidgwick 1907 , Nagel 1970 , Foot 1972 , Parfit 1984).
Once we have accepted the substantivists' claim that someone can be rationally criticizable for failing to care about her own future good, we might accept that someone can be rationally criticizable if she fails to care at all about the good of others. Indeed, substantivists typically hold that rationality not only rules out certain desires and patterns of desire but also requires certain desires, such as concerns for one's own future good and for the good of others.

6. Replies to Brandt's Proceduralism

How could substantivists respond to Brandt's proceduralism? The first thing they might say is that Brandt is wrong to assume that our desires necessarily become more rational as we obtain more empirical information. To take the most familiar example, suppose Bettina has a slightly below average natural concern for others. As we give her more information about the daily lives of the starving people in the world, her natural concern for them grows. Helping her to appreciate vividly their daily struggles, we have increased her natural sympathy.
But now suppose Bettina goes to work for aid agencies in the very worst hit areas. This puts her face to face with the pain, panic, and deprivation that so many millions suffer every day. At first, her natural concern for the worst off grows as she sees more directly and vividly how needy they are. But eventually the prolonged exposure to suffering all around Bettina thickens the skin over her heart. Eventually, even the desperate needs of others begin not to trouble her.

This story illustrates that, at least up to a point, some desires will intensify as we obtain further relevant information. But beyond that point, receiving yet more relevant information may serve to dampen or even extinguish these desires. Therefore, substantivists could claim, the best set of desires may not be the one we would have after exposure to maximum relevant empirical information. Too
end p.69

much information, even if relevant and true, can be overwhelming, even deadening (Gibbard 1990 , 165–66, 171–72, 175–77).
The other way that substantivists could reply to Brandt's proceduralism would be to appeal to counterexamples like the ones cited earlier. Consider a noninstrumental desire for a saucer of mud, or a noninstrumental desire to turn on radios whenever possible, or future Tuesday indifference, or within-a-mile altruism, or someone's complete lack of concern for his own future well-being, or someone's complete lack of concern for anyone's well-being other than his own. Suppose that one or more of these survives Brandt's "cognitive psychotherapy." Intuitively, it would still seem irrational to have these desires or patterns of concern.

7. Replies to Williams's Arguments for Proceduralism

How could substantivists respond to Williams's arguments for proceduralism? Recall that Williams's first argument, which appealed to explanation, made use of the premises:

(P3) An agent can be motivated by this reason to do this thing only if this agent either already believes that he or she has this reason or can come to believe that he or she has this reason by rational deliberation.
(P4) All reasons that an agent either already believes he or she has or can come to believe he or she has by rational deliberation are internal reasons.

In response to this argument, Parfit points out that for (P3) and (P4) to be true, "rational deliberation" must mean *procedurally* rational deliberation. However, Parfit claims, if someone has an external reason to do something, then if this agent "were substantively rational, his awareness of this external reason would motivate him" (Parfit 1997 , 116). So if we take "rational deliberation" to mean *substantively* rational deliberation, (C) does not follow.
Williams's second argument, which concerned the content of claims about reasons, made use of the premise:

(P2*)The content of the claim that there is an external reason for an agent to φ cannot be, and cannot entail that, the agent would be motivated to φ if he or she deliberated rationally.

end p.70

Hooker and Parfit point out that, again, for this premise to be true, "rational deliberation" must mean *procedurally* rational deliberation (Hooker 1987 , Parfit 1997). If we take "rational deliberation" to mean *substantively* rational deliberation, as substantivists do, this premise is false. And, in that case, (C*) does not follow.

How could substantivists respond to Williams's criticism of the view that the claim that there is an external reason for an agent to φ means that this agent would be nicer, more considerate, more courageous and the like if he or she were to φ?

As we have seen, Williams's criticism of this view is that, on this view, claims about external reasons are not saying anything distinctive. In response to this claim, substantivists could start by admitting that what it *means* for an agent to have an external reason to φ is not that this agent would be nicer, more considerate, more courageous and the like if he or she were to φ. Nevertheless, they could say, claims about niceness, considerateness and courage are in part ways of saying that an agent has an external reason to do something. For such claims normally *imply* that an agent has a reason to do the thing that these claims pick out as the nice, considerate, or courageous thing to do.

Alternatively, substantivists could reply that, for Williams's criticism to work, he must be assuming two things. First, he must be assuming that the meaning of claims about external reasons is exhausted by their truth-conditions. Second, he must be assuming that, on the substantivists' view, claims about niceness, considerateness, courage and the like have the same truth-conditions as claims about external reasons. If Williams is assuming the conjunction of these two claims, substantivists can reply in two ways. They can either:

(a)Deny that the meaning of a claim about external reasons is exhausted by its truth-conditions, [15]

or:

(b)Admit that the meaning of a claim about external reasons is exhausted by its truth-conditions, and say that the claim that there is an external reason for an agent to φ is true if and only if there actually is an external reason for this agent to φ.

If substantivists give reply (a), it is hard to see how Williams could still insist that claims about external reasons collapse into claims about niceness, considerateness, courage, and the like. Against (b), however, Williams might object that the truth-conditions that

substantivists propose here are vacuous. But substantivists could reply that this merely seems to be so, because the concept of a "reason" is a basic concept that cannot be analyzed in other terms. Moreover, they could say, if Williams assumes that the meaning of a claim about reasons is exhausted by its truth-conditions, he is himself guilty of collapsing the meaning of the claim that there is a reason for an agent to φ into the claim that, were the agent empirically well informed and reasoned logically from her existing set of motivations, the agent would be motivated to φ.

How could substantivists reply to Williams's criticism of the view that the claim that there is an external reason for an agent to φ means that if this agent were a well-informed and well-disposed deliberator, he or she would be motivated in these circumstances to do this thing?

As we have seen, Williams's criticism of this view was that, when the agent lacks the dispositions and capacities of a well-informed and well-disposed deliberator, what the agent has good reason to do cannot be determined by what the well-informed and well-disposed deliberator would do in the circumstances. In response to this criticism, substantivists can agree that any sensible view about practical reasons will hold that there is some connection between an agent's dispositions and capacities and what the agent has reason to do. Substantivists will typically accept that, *in many cases,* which reasons an agent has is related to which desires the agent has (Parfit 1997 , 128). For example, if Harry has a strong desire for food, he will normally have a reason to get some food. On a substantivist view, this reason is not *given by* Harry's desire, but is instead given by the fact that he will get pleasure from eating or by the fact that if he does not satisfy his hunger he will be too distracted to concentrate on anything else. Desire can influence what reasons agents have, according to substantivists, because of desire's pervasive connections with pleasure, concentration, and the like.

8. Rationality and Reasons

Brandt, Williams, Harman, and Parfit take *practical rationality* and *responding to reasons for action* to be very closely related. For Brandt, Williams, and Harman, practical rationality is primary, and what there is reason to do depends on what it is practically rational to do. Because Brandt, Williams, and Harman are proceduralists about practical rationality, they are also proceduralists about reasons for action. Hooker, Parfit, and others, by contrast, take responding to reasons for action to be primary. They take what it is practically rational to do to depend on what there are reasons to do. Since they are substantivists about reasons for action, they are substantivists about practical rationality.

A possible way out of this controversy is offered by T. M. Scanlon. According to Scanlon, it is not necessarily the case that if someone does not do what he has most reason to do, he or she fails to be fully practically rational. Instead, Scanlon claims, a failure of practical rationality occurs only "when a person recognizes

something as a reason but fails to be affected by it in one of the relevant ways" (Scanlon 1998 , 25). If we follow Scanlon in this, we can be proceduralists about practical rationality and substantivists about good practical reasons. Taking this position would enable us to avoid awkwardnesses in the two opposing positions.

Suppose that Steve is a cool, calculating, self-disciplined, efficient achiever of things that he wants and thinks important. His attitudes and actions conform very well to his own judgments about what to care about and pursue. Steve knows that he could reduce the suffering of innocent people massively merely by pushing a button that is right in front of him, which would cost him nothing beyond a millisecond of time. But he has no desire at all to push this button, and he would not reach such a desire by engaging in procedurally rational deliberation—because, even if he were presented with vivid empirical information about the suffering of these people, he would not care at all about their plight.

In that case, proceduralists about reasons will have to say that *there is no reason* for Steve to push the button. But that seems an awkward thing to say, because the thing that Steve could achieve is so important, and because he could achieve it with so little effort. Substantivists about practical rationality, on the other hand, will have to say that Steve *fails to be rational* in not pushing the button. But that seems an awkward thing to say as well, because Steve is such a cool, calculating, self-disciplined, efficient achiever of things that he wants and thinks important. Herein lies the appeal of Scanlon's compromise. If we are proceduralists about practical rationality, we can say that what is wrong with Steve is not that he fails to be rational. But if we are simultaneously substantivists about reasons for action, we can still say that there is something wrong with Steve, namely that there is a very strong reason for him to push the button, which he fails to see is a reason.

NOTES

Many thanks to Maria Alvarez, Piers Rawling, Mike Ridge, Philip Stratton-Lake, and Crystal Thorpe for very helpful comments on an earlier draft of this chapter.

1. We deliberately leave open here what it means to "rationally reach" a desire from one's present desires, since different proceduralists have different views on this, as will become clear below.

2. The "beliefs" we talk about in this section should be taken to be non-normative beliefs.

3. Philosophers who defend proceduralism will generally also hold that an agent can be open to rational criticism for *having* (rather than lacking) a desire only if the agent can rationally reach a state in which he or she *lacks* (rather than has) this desire from the beliefs and other desires that he or she has. To save words, in what follows, we will ignore this complication.

end p.73

4. To say that a person is open to rational criticism is not to say that this person is *irrational,* since a person can properly be called irrational only if he or she is open to severe rational criticism (see Parfit 1984 , 119; and Scanlon 1998 , 25–30).

5. More exactly, it is often called “Humeanism about normative reasons,” to distinguish it from Humeanism about motivating reasons (see, e.g., Smith 1994). Our focus in this chapter is on normative reasons.
6. We say “a new *empirical* belief” to exclude evaluative or normative beliefs, such as beliefs about what an agent has reason to do.
7. Williams himself does not call his conception of practical rationality “procedural.” This term is applied to Williams's view by Parfit 1997 .
8. It may be thought that (P3) should be formulated without “by rational deliberation.” But, given that Williams's conclusion is that all practical reasons are internal, and given how Williams defines internal reasons, he either is committed to (P3) as we have formulated it, or his conclusion does not follow.
9. Here and in what follows, “φ” represents the performance of an action.
10. We draw here on Harman 1977 . In essentials, Harman's views have not changed, as is apparent in Harman 1996 .
11. For Harman's endorsement of this, see Harman 1977 , 87, 125–28.
12. In Williams 1985 , Williams attacks “the morality system” (chap. 10). But Williams has since admitted an important place for judgments like “φ-ing is wrong” (Williams 1995c , 19–34, 32).
13. This is a leitmotiv of Harman's work in ethics. See, for example, Harman 1975 , Harman 1977 , 84, 106, and Harman 1996 .
14. To hold this is compatible with holding that one's own future well-being is less important than some other things, such as the well-being of others.
15. This will be most clearly true on an expressivist view about the meaning of judgments about reasons for action, such as R. M. Hare's view that the meaning of normative judgments (e.g., “A has reason to φ”) does not determine the truth conditions of such judgments (Hare 1981 , esp. 207).
end p.74

Chapter 5
HUMEAN RATIONALITY
Michael Smith

Hume famously thought that the scope of human reason was far more limited than many of us are inclined to think. As he puts it at one point, in an often quoted passage:
'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and have a more ardent affection for the former than the latter.In short, a passion must be accompany'd with some false judgement, in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgement. (Hume 1739 , 416)
In his discussion of this passage Mark Johnston suggests that “an adequate response might be: not contrary to reason in one sense perhaps, but brutally insane, psychopathically callous and demonically indifferent” (Johnston 1989 , 161). He goes on

to develop an alternative sense of the term “reason” according to which the preferences Hume cites would indeed count as contrary to reason. But while Johnston's response seems to me admirably to capture the reactions many of us have when we first read the passage just quoted, and though I agree with him that we need to develop alternative senses of the terms “reason” and, for that matter, of its cousin “rationality,” it seems to me that we must face up to the fact that the view Hume puts forward in this passage is difficult to avoid. Hume gives a plausible argument for the claim that the terms “reason” and “rationality” have just one sense, a sense that
end p.75

makes what he says seem inevitable (see also Hooker and Streumer, chap. 4, O'Neill, chap. 6, and McNaughton and Rawling, chap. 7, this volume).

In what follows I will begin by explaining how, as I understand things, Hume is led to his (grotesque) conclusion. The explanation lies in his view that the concepts of reason and rationality are best explained by reference to their relations in the theoretical domain, specifically in the domain of deductive reasoning. As we will see, this leads Hume to take a very radical view about the scope of reason and rationality even in other aspects of the theoretical domain—namely in the domain of inductive reasoning. I will then consider how we might avoid Hume's conclusion. The issue, to anticipate, is whether, once we liberalize our understandings of the terms “reason” and “rationality” in the way required to take a more sensible view about the nature of reason and rationality in the theoretical domain, there is a stable position left to take in the practical domain that retains anything of the spirit of Hume's remarks. The question, in other words, is whether a sensible liberalizing of Hume on the nature of reasons and rationality sends us down a slippery slope all the way to Kant.

Before starting the discussion proper, however, let me offer the following disclaimer. Though in what follows I will speak incautiously about what Hume's views are about the nature of reasons and rationality, my real interest lies not in what Hume, the historical figure, thought about these matters. Indeed, I have become convinced that Hume's own views are far more complicated, and certainly far more controversial and beholden to the times at which he was writing, than they are normally taken to be in contemporary discussions (compare Stroud 1977 , Baier 1991 , Snare 1991 , Millgram 1995 , Bricke 1996 , and Owen 2000). What really interests me, then, is not what Hume himself thought about these matters, but rather, as the title of the chapter makes plain, what a Humean, a contemporary philosopher whose philosophical views have been greatly influenced by certain of Hume's writings, would have to say (see especially Davidson 1963 ; Williams 1980 , 1995a ; Gauthier 1986 ; Bratman 1987 , 1999 ; Mele 1987 , 1992 , 1995 ; Lewis 1988 ; Copp 1997 ; Dreier1997 ; Railton 1997 ; Blackburn 1998).

1. The Radical Humean View about the Relationship between Reasons and Rationality

Our topic is to be the relationship between reasons and rationality. Note that we can make a rough distinction between two domains in which these concepts have application: the theoretical domain and the practical domain. The theoretical domain is the realm of belief formation: that is, the realm in which we come to a view about the way the world is. The practical domain, by contrast, is the realm of desire formation: that is, the realm in which we become disposed to make the world be one way rather than another.

Hume's view, of course, was that the theoretical and practical domains are utterly distinct from each other. This is because, inter alia, belief and desire can always be pulled apart, at least modally. No matter what beliefs a subject has, and what desires, we can always imagine a possible world in which the subject has those beliefs but has different desires, and vice versa (Smith 1987 , 1988). Hume thus rejects the possibility of there being any beliefs that are desires (though contrast McDowell 1978): that is, to use James Altham's wonderful term, the possibility of besires (Altham 1986). It is also because, as Hume sees things, no belief can rationally produce a desire: “Thus it appears, that the principle, which opposes our passion, cannot be the same with reason, and is only call'd so in an improper sense. We speak not strictly and philosophically when we talk of the combat of passion and of reason. Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them” (Hume 1739 , 415). Note, however, that we haven't yet assumed that the theoretical and the practical domains are utterly distinct from each other in either of these ways. For all that we have said so far, these domains may well overlap. Whether or not they overlap, and if they do the extent to which they do, will emerge in what follows.

With the distinction between the theoretical and practical domains in place, let's now restrict our attention to the theoretical domain and ask, What is it to have a reason, and rationally to form beliefs on the basis of that reason, in that domain? As I understand it, Hume thinks that we should answer this question by generalizing from a case in which reasons and rationality stand in a more or less transparent relation, namely from the case of deductive reasoning.

Suppose I believe that *p*, and I believe that if *p* then *q*, and I rationally go on to form the belief that *q*. What should I say if I were to be asked what my reasons were for forming that belief? The answer would seem plain enough. I should say that my reasons for forming the belief that *q* were that *p* and that if *p* then *q*. Put another way, I should say that *p*, and if *p* then *q*, are the reasons why *q*. Of course, the fact that I should say this does not entail that there are, in fact, reasons why *q*. All that it entails is that, if things are as I take them to be, then there are reasons why *q:* namely that *p* and that if *p* then *q*. This, in turn, suggests an explanation of why the transition between my beliefs is a rational one, an explanation that presupposes the possibility of there being reasons why the things I believe are true. Suppose I begin by believing that *p*, and believing that if *p* then *q*, and, on the basis of these beliefs, come rationally to believe that *q*. The obvious explanation of the rational transition between my

beliefs is that, inter alia, there is an isomorphism between their relations and the logical relations between the propositions I believe, that is, the propositions which give the reasons why *q* (at least if things are as I take them to be).

This idea can be neatly captured in table 5.1 (compare Pettit and Smith 1990). What we have on the left hand side is a set of transitions between psychological states, my beliefs, and what we have on the right hand side is a set of relations between the propositions to which I would advert in giving expression to my various beliefs: that is, the propositions I believe. The suggestion just made is that the explanation of why a rational transition from the beliefs mentioned in (1) and (2) to the belief mentioned in (3) is possible—in other words, the explanation of why the beliefs mentioned in (1) and (2) may rationally give rise to the belief mentioned in (3)—is that the propositions mentioned in (1) and (2) logically entail the proposition mentioned in (3). The possibility of a rational transition between beliefs in this way seems to presuppose the possibility of there being independent reasons. The rationality of the transitions between the beliefs is derived from the logical relations between the propositions believed. (A good question to ask at this point is how beliefs are supposed rationally give rise to other beliefs. What is the mechanism by which this happens? There must be such a mechanism because, since irrationality is possible, the mere possession of the beliefs responsible for the rational production cannot be sufficient all by itself. I will, however, postpone answering this question until later.)

Table 5.1

Transitions among Psychological States	Relations between Propositions
(1) I believe that *p*	(1′) *p*
(2) I believe that if *p* then *q*	(2′) If *p* then *q*
So (3) I believe that *q*	Therefore (3′) *q*

It will, I hope, be admitted that this is an attractive, even if somewhat revisionary, account of the relationship between reasons and rationality in the case of deductive reasoning (see also Broome 1999 , 2001b). The idea that reasons are propositions that logically entail the propositions for which they are reasons would, after all, seem to comport well with our prereflective conception of a reason as a consideration that justifies (Smith 1987 , 1994 ; Dancy 2000). For the considerations that justify do indeed seem to be propositions. Moreover it also seems plausible to suppose that the rationality of forming a belief depends on our believing there to be reasons for doing so—or, more generally, that the rationality of a psychological transition must have something to do with the possibility of there being reasons for that psychological transition. To this extent, rationality does indeed seem to presuppose that we believe there to be reasons.

Having said that, however, it must also be admitted that the account is at least somewhat revisionary. For it entails that whenever we talk of having reasons for our beliefs, as we often do and indeed just did, we are at best speaking loosely. What we ordinarily describe as “reasons for beliefs” are really reasons why the propositions we believe are true: the reasons are true propositions that logically entail the propositions we believe. Strictly speaking we should therefore say that when people believe a proposition that, if it were true, would provide a reason why some other proposition they believe is true, then the first belief of theirs can make the second belief rational.

end p.78

Attractive though it might be, however, the generalization of this Humean account of the relationship between reasons and rationality beyond the realm of deductive reasoning has some rather disturbing consequences. There are, after all, many cases in which we would ordinarily take there to be (speaking loosely again) reasons for forming beliefs where the reasons in question do not have this character. Think of cases in which we have inductive evidence, such as those cases in which we form a belief as a result of inference to the best explanation.

Suppose, for example, I believe that the barometer is falling, and I believe that the best explanation of the barometer's falling is that something is happening that will cause it to rain tomorrow. Suppose further that, as a result, I go on to form the belief that it will rain tomorrow. For this to be a case of rational belief formation the transition between my psychological states would have to be one whose structure is isomorphic to a set of logical relations between the propositions believed. The problem is, however, that there is no such structure, as table 5.2 makes plain. The propositions mentioned in (4) and (5) do not satisfy the condition formulated above for their being reasons for the proposition mentioned in (6): that the barometer is falling, and that the best explanation of barometer's falling is that something is happening that will cause it to rain tomorrow, simply does not logically entail that it will rain tomorrow. Nor will it help to suggest that the evidence in any real case would be much richer than we have supposed it to be in this case. For the crucial point is that the hallmark of inductive reasons—reasons such as those provided by the consideration that something or other is the best explanation of some aspect of our experience—is precisely that they do not logically entail the conclusions that we think that they are reasons for.

Table 5.2

Transitions among Psychological States	Relations between Propositions
(4) I believe that the barometer is falling	(4′) The barometer is falling
(5) I believe that the best explanation of the barometer's falling is that something is happening that will cause it to rain tomorrow	(5′) The best explanation of the barometer's falling is that something is happening that will cause it to rain tomorrow
So (6) I believe that it will rain tomorrow	Therefore (6′) It will rain tomorrow

The upshot is thus that, if a condition on a set of propositions being reasons why some conclusion is true is that those propositions logically entail that conclusion, then it follows that there are no inductive reasons. Hume famously embraces this conclusion. As he puts it at one point, “Even after the observation of a frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience” (Hume 1739 , 139). But if the category of inductive

reasons is empty, and if the rationality of a transition between beliefs presupposes the possibility of reasons,
end p.79

then it follows that there is no such thing as rationally forming a belief on the basis of inductive evidence; no such thing as, for example, rationally inferring to the best explanation. The formation of beliefs on these bases is not rational. The beliefs mentioned in (4) and (5) cannot rationally give rise to the belief mentioned in (6).
Having seen the way in which Hume is forced to distinguish, within the theoretical domain, between those beliefs that do and those that do not come within the orbit of reasons and rationality, it should come as no surprise that he holds a radical view in the practical domain. Consider the example he gives at the beginning of the passage quoted at the outset.
Hume tells us that it is not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. In order to resist this conclusion consistently with the view of reasons and rationality we have derived from the deductive case, we can now see that we would have to suppose that its being the case that scratching my finger would cause the destruction of the whole world (supposing this to be the case) provides a reason for its being the case thata reason for its being the case that what? A moment's reflection on that question should make it plain that the practical case is even more problematic from the point of view of reason and rationality than the inductive case.
Table 5.3 makes the problem vivid. For there to be a rational transition from the belief mentioned in (7) to the preference mentioned in (8), we can now see, following Hume, there would have to be an isomorphism between the relations between this belief and preference and the logical relations between the propositions to which I would advert in giving expression to the belief and preference. But it seems that there is no proposition to which I would advert in giving expression to my preference not to scratch my finger. Propositions are what I advert to in giving expression to my beliefs, as in (7), not my preferences; hence the question marks expressing bewilderment in (8). But in that case it follows that there are no such logical relations, and hence no such isomorphism. If Hume is right that rationality presupposes the possibility of a logical relationship between
end p.80

independent reasons, then there is no such thing as a rational transition from the belief mentioned in (7) to the preference mentioned in (8).

Table 5.3

Transitions among Psychological States	Relations between Propositions
(7) I believe that scratching my finger would cause the destruction of the whole world	(7′) Scratching my finger would cause the destruction of the whole world
So (8) I prefer not to scratch my finger	Therefore (8′) ???

The practical case is thus even more deeply problematic than the inductive case. In the inductive case at least there are some propositions to which I would advert in giving expression to my beliefs: there are, in other words, considerations that purport to justify. The problem in that case is that, though there are such propositions, they don't stand in the required logical relations to each other. In the practical case, however, the problem is that there are no propositions to be so much as candidates for propositions that stand in the required logical relations. It will emerge presently that this difference between the two cases is of some significance.

It might be thought that we could avoid this conclusion if we were to suppose, as some have, that we advert to optatives or imperatives in giving expression to our preferences (Hare 1952 , Goldman 1970 , Castañeda 1975 , Kenny 1989). According to this view my preference not to scratch my finger is expressed in the claim "Would that it be the case that I not scratch my finger," or perhaps in the claim "Let it be the case that I do not scratch my finger." If these theorists were right, then the idea is we could develop a logic that would make it plain which logical relations obtain between claims of these kinds and more ordinary propositions. But attempts to develop a logic of optatives or imperatives have not been promising, leading some to conclude that there is no such logic (Harman 1986). Moreover, others have suggested that the best way in which to understand why optatives and imperatives sometimes seem to stand in logical relations, insofar as they do, is because they are really just disguised propositions (Lewis 1970). According to this view, "Let it be the case that I do not scratch my finger" means much the same as "I command myself to scratch my finger." But, of course, if this is right then it turns out that there is no special class of expression that expresses our preferences. For the proposition "I command myself to scratch my finger" is plainly an expression of my belief that I command myself to scratch my finger, not my preference to scratch my finger.

We can now see why Hume holds that the theoretical and the practical domains are utterly distinct from each other. For he has chosen, in the belief mentioned in (7) and the preference mentioned in (8), a belief and desire that are perhaps among the most promising candidates for a belief that is a desire (or a besire), if such there be; or, if there are none such, for a belief that is rationally related to a desire, again if such there be. But the belief mentioned in (7) and the desire mentioned in (8) are plainly distinct existences, for we have no difficulty in imagining someone who has the belief without the desire, and vice versa. We simply have to imagine people who are, in Johnston's terms, "brutally insane, psychopathically callous and demonically indifferent." And, if the argument given above is sound, then Hume has in effect demonstrated that the belief and desire cannot be rationally related either. For there is no candidate for one of the propositions that would have to stand in the required logical relation.

Let's recap the argument of this section. Hume offers an account of the relationship between reasons and rationality according to which the only realm in which there are reasons and rational transitions between psychological states is the realm of deductive reasoning. There are no reasons or rational transitions between psychological states in the realm of inductive reasoning, nor are there reasons or rational transitions between psychological states in the realm of practical reasoning. The expressions "inductive reasoning" and "practical reasoning" thus turn out to be oxymorons. This is plainly a radical view. Radical though it is, however, I hope it has emerged that the view is at least

based on an argument and that that argument has a certain attraction. If we do not like Hume's conclusions, as I assume we do not, then we have no choice but to say where, as we see things, his arguments for those conclusions go wrong.

2. A More Moderate, but Still Quite Humean, View about the Relationship between Reasons and Rationality

Unsurprisingly, many theorists have found it extremely difficult to accept Hume's radical view about the relationship between reasons and rationality. But how are we to resist the radical view? How should we go about constructing a more moderate position?
By all accounts, in order to resist Hume's view we must give up on the strategy of explaining the rationality of transitions between psychological states by reference to the logical relations between the propositions to which we would advert in giving expression to those psychological states: that is, by reference to reasons, in the sense gleaned from the case of deductive reasoning. Instead we must focus on the transitions between the psychological states themselves and come up with
end p.82

an alternative explanation of why they constitute rational transitions, an account that is independent of the logical relations between the propositions, if such there be, to which we would advert in giving expression to those psychological states. However, in so doing we should try to remain faithful to the idea that the rationality of forming a belief has something to do with the possibility of there being reasons for forming that belief, for that idea has its own independent appeal.
It is not difficult to see how we might do this, at least in broad outline (Strawson 1952). Suppose that someone believes both that the barometer is falling and that the best explanation of the barometer's falling is that something is happening that will cause it to rain tomorrow, and on the basis of these beliefs goes on to form the belief that it will rain tomorrow. In that case it is very plausible to suppose, at least prereflectively, that there is, to that extent at least, far more coherence in that set of beliefs than there is in the set of beliefs that the subject would have had if he hadn't gone on to form the belief that it will rain tomorrow. A subject who has the beliefs mentioned in (4), (5), and (6) has, to that extent, a more coherent belief set than one who has the beliefs mentioned in (4) and (5) but lacks the belief mentioned in (6). But, if this is right, then we might conclude that the transition is, for this very reason, a rational transition. Indeed, we might even conclude that the propositions believed constitute reasons for forming the belief that it will rain tomorrow: that is, that the propositions mentioned in (4) and (5) are considerations that justify the formation of the belief mentioned in (5). What is crucially important, however, is that the order of explanation is the reverse of that proposed by Hume. What comes first is the explanation of the rationality of the psychological transitions between the beliefs: that is, the claim about the coherence of the belief set. The claim about the propositions believed being reasons is derivative from that fact and has no independent force.

Of course, in order to tell this story in a fully convincing way we would have to make good on the prereflective conception of coherence. The coherence of a belief set would have to require more than that the beliefs be logically consistent, for example. We must therefore have in mind something more like explanatory coherence, or probabilistic coherence, or something else along these lines. However, for present purposes it isn't necessary for us to go into detail about the conception of coherence (though see Harman 1986). It will suffice if the following claims seem plausible: first, that a set of beliefs formed rationally on the basis of inductive evidence does indeed display a distinctive kind of coherence, and second, that we could, at least in principle, give a descriptive account of what that kind of coherence consists in. In other words, all that is important is that it be agreed that the story of reason and rationality can begin with an independent story of rationality so that the story of reasons can be derived from it, rather than vice versa.

If some such story can be told about the nature of rational belief formation in the case of inference to the best explanation, an obvious question presents

itself: Can we tell a similar story in the practical realm? In other words, can we give a similar account of the transitions between desires and beliefs that constitute rational transitions? Unsurprisingly, perhaps, it is not clear that we can. A further difficulty presents itself in the practical realm, one related to the difficulty mentioned earlier when we saw that the practical case is even worse than the inductive case.

In order to see this difficulty, it will help if we focus on a transition between desires and beliefs that nearly everyone will agree is rational. Consider cases of means-end rationality, that is, cases in which I form a desire to perform the means to an end because I have a background desire for that end and a belief that the means is a means to that end. Suppose, for example, that I desire to relieve an itch and believe that I can do so by scratching my finger, and, on this basis, go on to form the desire to scratch my finger. As I said, most of us would suppose this to be rational transition. But can we tell a story similar to the story that we just told in the case of inference to the best explanation in order to explain why it is rational?

Table 5.4 illustrates the difficulty. To be sure, it does indeed seem, at least prereflectively, as though someone who has the desire mentioned in (9) and the belief mentioned in (10), and who on this basis goes on to derive the desire mentioned in (11), has a much more coherent set of desires and beliefs than someone who has the desire mentioned in (9) and the belief mentioned in (10), but who doesn't go on to derive the desire mentioned in (11). Moreover it might well be thought that this suffices to explain why we suppose that having the desire mentioned in (9) and the belief mentioned in (10) makes it rational to have the desire mentioned in (11). To this extent it might look like (9), (10), and (11) are straightforwardly analogous to (4), (5), and (6).

Table 5.4

Transitions among Psychological States	Relations between Propositions

(9) I desire that I relieve the itch in my finger	(9′) ???
(10) I believe that I can relieve the itch in my finger by scratching it	(10′) I can relieve the itch in my finger by scratching it
So (11) I desire that I scratch my finger	Therefore (11′) ???

The difficulty, however, is that there remains a significant disanalogy between the two cases. Since there are no propositions to which we would advert in giving expression to the desires mentioned in (9) and (11)—no candidates for propositions to be mentioned in (9) and (11), unlike (4), (5), and (6)—it follows that we cannot derive from this account of the rationality of forming the desire mentioned in (11) an account of the reasons we have for forming that desire. The disanalogy, in other words, is that we cannot square the idea that this is a rational transition with the independently attractive idea that the rationality of a psychological transition must have something to do with the possibility of there being reasons for that psychological transition. The upshot is that, since there are no means-end reasons, there can be no such thing as means-end reasoning or means-end rationality (Millgram 1995).

This conclusion is likely to come as a surprise to many Humeans, for they are inclined to say that, in the example described above, my reason for forming the desire to scratch my finger is that I desire to relieve the itch in my finger and can do so by scratching it (see, for example, Williams 1980 ; Gauthier 1986 ; Brat

man 1987 , 1999 ; Dreier 1997). But, as I hope the table above makes plain, Humeans who say this sort of thing are plainly confusing the fact that there is a desire that one has when one engages in so-called means-end reasoning—namely in our example, the desire that I relieve the itch in my finger mentioned in (9)—with the falsehood that one would advert to a proposition to the effect that one has that very desire in giving expression to that desire: that is, in our example, with the idea that the proposition that I desire to relieve the itch in my finger should figure in (9). That is, they confuse the fact that I have a desire for an end, which is a psychological fact about me, with the claim that my possession of that desire is a consideration that justifies.

In order to see that that is a confusion, it suffices that we ask ourselves the following question. What proposition would have figured in (9) if the psychological state mentioned in (9) hadn't been the desire that I relieve the itch in my finger, but had instead been the belief that I desire to relieve the itch in my finger? The answer to this question is obvious. The proposition that I desire to relieve the itch in my finger would in that case have figured in (9). But that proposition can hardly be the proposition to which we would advert in expressing both my desire to relieve the itch in my finger and the belief that I desire to relieve the itch in my finger! But since it is plainly the proposition to which we would advert in expressing the belief, the only conclusion to draw is that it isn't the proposition to which we would advert in expressing the desire (Pettit and Smith 1990). We must therefore conclude that those Humeans who think that the fact that I desire to relieve the itch in my finger is a reason to which we can appeal in so-called means-end

reasoning—a consideration that justifies, in other words—are simply mistaken. There are no considerations that justify, or, at any rate, none corresponding to the desires.
Let's recap. We can resist Hume's radical view that there are no inductive reasons, and hence the view that there is no such thing as rationally forming a belief on the basis of such reasons, if we suppose that we have an independent grip on what would constitute a rational transition between beliefs about inductive evidence and those beliefs that we go on to form on the basis of that evidence.
end p.85

The idea that such beliefs display a distinctive kind of coherence looks like it could provide us with that independent grip. Such transitions are rational when they display the appropriate kind of coherence. This allows us to remain faithful to the idea that the rationality of a psychological transition must have something to do with the possibility of there being reasons for that psychological transition, for we can then suppose that the idea of an inductive reason is simply derived from the account of inductive rationality. The considerations that justify are those to which we would advert in giving expression to the beliefs that stand in the appropriate kind of coherence relations. In this way we restore the possibility of inductive reasons and reasoning.
Unfortunately, however, no similar story can be told about rationally forming desires in the practical realm. For even though there do appear to be rational transitions from (say) desires for ends and beliefs about means to desires for means, we cannot derive from this a parallel story about the reasons we have for forming those desires for means. In other words, in the practical case we cannot remain faithful to the idea that the rationality of a psychological transition must have something to do with the possibility of there being reasons for making that psychological transition. We must therefore conclude that the appearance of rationality in the practical case is, much as Hume said, an illusion. There is no such thing as means-end rationality: talk of "means-end reasoning" remains an oxymoron. There is simply the human habit of forming desires for means on the basis of desires for ends and beliefs about means, a habit that is underwritten by neither reasons nor rationality.

3. An Even More Moderate, but Still Somewhat Humean, View about the Relationship between Reasons and Rationality

Many of us will find it hard to believe that there is no such thing as means-end rationality. But how would we go about constructing an even more moderate position than the moderate Humean position just described, the one that allows for deductive and inductive reasons and rationality but not for practical reasons and rationality, not even means-end reasons and rationality? The answer should be obvious.
Up until now we have simply been assuming that there must be some sort of nexus between reasons and rationality. More specifically, we have been assuming that the rationality of a psychological transition must have something to do with the possibility of there being reasons for that psychological transition. But the obvious conclusion to draw

in the light of our discussion of means-end rationality is that there need be no such nexus. Nor would it be ad hoc to draw this conclusion in the light of that discussion. For, as we saw, people who fail to desire the believed means to their desired ends do not just seem to be unusual, do not just seem to lack a typical human habit. Their psychologies seem rather to suffer from a distinctive kind of incoherence, an incoherence that would be removed if they acquired a desire for the believed means to their desired ends. Since it appears that that kind of incoherence need have nothing to do with being insensitive to reasons, we should take that appearance of incoherence at face value and conclude that they are indeed irrational: means-end irrational. Of course, this would require us to further extend the story of coherence in some appropriate way. But again, much as with the case of inductive rationality, it seems extremely plausible to suppose that that could be done (Harman 1986).

If all this is agreed, then the upshot is that there is an even more moderate Humean position available. According to this even more moderate view, rational transitions are possible wherever we find psychological states that display the appropriate kind of coherence relations. Since these relations are found in both the theoretical and the practical domains, it follows that there is both theoretical and practical rationality. There is, however, the following difference between the two domains. In the theoretical domain there are propositions we believe that justify our acquiring the beliefs we acquire. It therefore follows that, in this domain, there aren't just rational transitions, but that there are also reasons and reasoning. But because, in the practical realm, there are simply desires for ends and beliefs about means to those ends that stand in certain relations of coherence to desires for those means, it follows that in the practical realm there are just the rational transitions themselves. There are no means-end reasons and no means-end reasoning.

Note that we are now in a position to answer a question we postponed earlier. How are beliefs supposed rationally to give rise to other beliefs? More generally, how are psychological states supposed rationally to give rise to other psychological states? What is the mechanism by which this happens? Given what has just been said, the obvious answer is: in virtue of the fact that the subject of those psychological states is rational, where the subject's being rational is a matter of her having, and exercising, a capacity to have a coherent set of psychological states (Smith 1997 , 2001). Moreover, and importantly, since this capacity has to explain rational production across the board—that is, since it must explain both how beliefs rationally produce other beliefs, and how desires for ends and beliefs about means rationally produce desires for means—this capacity cannot in turn be thought of on the model of a desire for an end, such as the end of having a

coherent psychology, which works by combining with a belief about the means by which this is to be achieved to rationally produce a desire for that means. For how, after all, would we explain that instance of rational production (Dreier 1997 , Smith 2001)? Instead we must conceive of the capacity on the model of the inferential dispositions Lewis Carroll taught us all to believe in (Carroll 1895).

4. The Instability of the Even More Moderate, but Still Somewhat Humean, View about the Relationship between Reasons and Rationality

There is, however, a problem with the even more moderate Humean view. Once we admit that desires can stand in rational relations to each other—coherence relations—it follows, more or less immediately, that we can construct the idea of a consideration that justifies our acquiring a certain desire. And once we have constructed the idea of a consideration that justifies our acquiring certain desires, it is hard to see in what sense we are any longer committed to a Humean view about the relationship between reasons and rationality at all. We seem to have crossed over a threshold. Let me explain.

Many have recently argued that some version of the dispositional theory of value looks to be extremely plausible (Firth 1952 ; Brandt 1979 ; Johnston 1989 ; Lewis 1989 ; Smith 1989 , 1994 , 1997). According to the version of the theory that I myself favor, for example, when a subject judges it desirable that *p* be the case in certain circumstances *C*, what she has is a belief that she would want that *p* be the case in *C* if she had a fully rational desire set—or, more generally perhaps, if she had a set of desires that eluded all forms of rational criticism. Facts about desirability are, in this way, constructed from facts about the desires we would have if we were fully rational. As we will see, it is this fact about desirability—that is, the fact that something is what a subject would desire if she had a desire set that eluded all forms of rational criticism—that looks like it constitutes a consideration that can justify that subject in acquiring a corresponding desire. In order to see that this is so, however, we need first to spell out the dispositional theory more fully. In particular, we need to say what would make it the case that a desire set eludes all forms of rational criticism.

Following the lead of Bernard Williams, who is in turn inspired by something that Hume says in the passage quoted at the outset—"a passion must be accom

pany'd with some false judgement, in order to its being unreasonable"—we might begin with the observation that the fact that a subject's desires are based on ignorance or error itself looks to be grounds for rational criticism of that desire (Williams 1980). This is because someone who was perfect, from the point of view of reason, would plausibly be omniscient and make no mistakes. This suggests that someone who (say) desires to relieve an itch and who believes, falsely, that she can relieve her itch by staring at her finger, and who on this basis goes on to form the desire to stare at her finger, fails for this very reason to have a desire set that eludes all forms of rational criticism. For a desire set that eludes all forms of rational criticism would contain no desires based, as her desire to stare at her finger is, on a false belief. Likewise, someone who desires to relieve an itch and who can relieve her itch by scratching her finger, but who doesn't believe that this is so, fails to have a desire set that eludes all forms of rational criticism as well. For someone who has such a desire set would lack no desires due simply to her ignorance. Moreover there are other ways in which a subject's desires might become vulnerable to rational criticism too. For example, as we saw in the earlier discussion of means-end

rationality, to say that a subject has a desire set that, as a whole, exhibits incoherence is plainly a rational criticism of that desire set. Nor should it be thought that cases of means-end irrationality are the only cases in which we find incoherence in a desire set. There would seem to be other such cases as well (see, for example, Smith 1994 , 1997 ; for an alternative view, see Sayre-McCord 1997).

For example, imagine that I desire that person *A* fares well, and I desire that person *B* fares well, and I desire that person *C* fares well, and so on, but that I don't desire that person *Z* fares well. Suppose further that when asked why I don't desire that *Z* fares well I can't identify any feature of *Z* that distinguishes him from *A*, *B*, *C*, and the rest—apart, of course, from the fact that he is *Z*. Perhaps I don't discriminate between *Z* and the others for any other purpose except his faring well. It is surely then plausible that I could quite rightly thereby come to see my lack of a desire that *Z* fares well as completely arbitrary. But if my lacking a desire that *Z* fares well is completely arbitrary, then it surely makes perfect sense to say that my desire set suffers from a corresponding kind of incoherence. We can represent the situation in table 5.5 . The suggestion is that there is incoherence in a psychology that includes the elements mentioned in (12) and (13) but does not include the desire mentioned in (14). The incoherence lies in the fact that that pattern of desire and indifference doesn't fit well with the belief that an arbitrary distinction is being made. In this case too, then, it seems that my desire set would be vulnerable to rational criticism, criticism that I could avoid only by acquiring the general desire that people fare well and, on that basis, together with my belief that *Z* is a person mentioned in (15), acquiring the desire that *Z* fares well mentioned in (16) as well.

end p.89

Table 5.5

Transitions among Psychological States	Relations between Propositions
(12) I desire that person *A* fares well, I desire that person *B* fares well, I desire that person *C* fares well,and I desire that person *Y* fares well.	(12′) ???
(13) I believe that it is arbitrary to distinguish person *Z* from *A–Y*	(13′) It is arbitrary to distinguish person *Z* from *A–Y*
So (14) I desire that people fare well	Therefore (14′) ???
(15) I believe that *Z* is a person	(15′) *Z* is a person
So (16) I desire that *Z* fares well	Therefore (16′) ???

Now that we have spelled out a little of what it means to say that a desire set eludes all forms of rational criticism, the crucial point to note is that this fact could be the object of a subject's belief (Smith 1994). Thus, imagine a subject who comes to believe that (say) she would desire that she scratches her finger in the circumstances of action that she presently faces if she had a maximally informed and coherent desire set, but suppose further that she doesn't have any desire at all to do so. Now consider the pair of

psychological states that comprises her belief that she would desire that she scratches her finger in the circumstances of action that she presently faces if she had a maximally informed and coherent desire set, and which also comprises the desire that she scratches her finger, and compare this pair of psychological states with the pair that comprises her belief that she would desire that she scratches her finger in the circumstances of action that she presently faces if she had a maximally informed and coherent desire set, but which also comprises instead indifference to scratching her finger, or perhaps an aversion to her doing so. Which of these pairs of psychological states is more coherent?
The answer would seem to be plain enough (see Smith 1994 , 1997 , 2001 ; for an alternative view, see Sayre-McCord 1997 , Schafter-Landau 1999). The first pair is much more coherent than the second. There is disequilibrium or dissonance or failure of fit involved in believing that you would desire yourself to act in a certain way in certain circumstances if you had a maximally informed and coherent desire set, and yet not desiring to act in that way. The failure to desire to act in that way is, after all, something that you yourself disown; from your perspective it makes no sense, given the rest of your desires; by your own lights it is a state that you would not be in if you were in various ways better than you actually are: more informed and more coherent in your desiderative outlook. There would therefore seem to be more than a passing family resemblance be

tween the relation that holds between the first pair of psychological states and more familiar examples of coherence relations that hold between psychological states.
If this is right, however, then the upshot is that table 5.6 , too, is a rational transition between psychological states. This transition is a rational one because it too is underwritten by coherence. Moreover, and more importantly, it even looks plausible to suppose that the proposition mentioned in (17)—that is, that I would desire that *p* in *C* if I had a maximally informed and coherent desire set—is a consideration that, if true, justifies my forming the desire that *p*. What better justification could there be for me to form the desire that *p?*

Table 5.6

Transitions among Psychological States	Relations between Propositions
(17) I believe that I would desire that *p* in *C* if I had a maximally informed and coherent desire set	(17′) I would desire that *p* in *C* if I had a maximally informed and coherent desire set
So (18) I desire that *p*	Therefore (18′) ???

If this is agreed, however, then we might well begin to wonder whether we any longer have any reason to suppose, as Hume did, that the theoretical and the practical domains are utterly distinct from each other. The problem is not that we have come to think that there are beliefs and desires that cannot be pulled apart from each other modally. That possibility has been granted in all of our discussions. The problem is rather that, granting

that possibility, we now seem to have an example of a belief that can rationally require the acquisition of a desire all by itself, something Hume had claimed to be impossible. We have, in other words, a case in which reason can produce a motive. For the belief mentioned in (17) can, if the subject is rational—that is, if it operates in conjunction with the subject's capacity to have a coherent psychological state (see again the discussion at the end of the last section)—produce the desire mentioned in (18) without the aid of any desire.

Indeed, we might even begin to wonder whether we shouldn't reconsider the psychological transition from (7) to (8) with which we began—that is, the transition from the belief that scratching my finger would cause the destruction of the whole world to the preference not to scratch my finger. For what we have, in effect, demonstrated is that Hume's argument for supposing that this isn't a rational transition is completely fallacious (see also Korsgaard 1986). It is irrelevant that there is no proposition to be mentioned in (8), and hence that there is no candidate proposition to be logically entailed by the proposition mentioned in (7). The rationality of a psychological transition requires no such thing, as we have learned from our discussion of means-end rationality. The crucial question is rather whether the pairing of the belief that scratching my finger would cause the destruction of the whole world and the preference to scratch my finger is an especially coherent pairing, as compared with the pairing of that belief together with indifference to scratching my finger, or an aversion to scratching my finger.

The truth is that I don't myself see how to answer that question decisively one way or the other. On the one hand, it is tempting to simply recall Mark Johnston's remark, quoted at the outset, that though the pairing of the belief with indifference or aversion is "not contrary to reason in one sense perhaps," it is "brutally insane, psychopathically callous and demonically indifferent," and to insist that the conceptions of insanity and psychopathy that Johnston was quite rightly drawing on here are plainly predicated on the supposition of incoherence. However, on the other hand, we all know that such normative conceptions of insanity and psychopathy are themselves hotly contested (Szasz 1961). What we really need here is an independent argument that would clinch the case one way or the other.

As I understand it, it is an independent argument of this kind that Kantians have been searching for all along (Kant 1786 , Nagel 1970 , Korsgaard 1996b). Humeans have always insisted that the Kantians' search for such an argument is in vain, because Hume provided a decisive refutation of the very possibility of there being such an argument. But, if what we have said here is right, then both Hume and the Humeans are wrong to suppose that he provided a decisive refutation of any such thing. Humeans and Kantians alike should therefore turn their attention to the arguments the Kantians come up with and evaluate them on a case-by-case basis. Whether the arguments provided here have put us on slippery slope all the way to a Kantian view of reasons and rationality is yet to be determined.

end p.92

Chapter 6 KANT
Rationality as Practical Reason

Onora O'Neill

Kant is famous for undertaking a critique of reason and for calling two of his most significant works *critiques of reason*. [1] These titles raise suspicions. Does Kant genuinely criticize reason, thereby calling into question the very processes by which any reasoned thought or action—including any criticism of reason—should be conducted? Or does he give these pretentious titles to works that deploy rather than criticize reason? Indeed, could anything really, seriously count either as a *critique* of reason or as a *vindication* of reason? Isn't the very idea that we could *show* that certain ways of thinking or acting are reasoned or reasonable absurd? After all, the demonstration must either build on assumptions that lack reasoned vindication or be supported by arguments that deploy the very conception of reason supposedly vindicated. So it will be either unreasoned or circular: either way it will fail to vindicate reason. We have grounds for suspecting that no ways of organizing thinking or acting have unconditional authority, and that Kant *cannot* have vindicated reason.

Kant's attempt to give an account of practical reason that offers unconditional reasons for action and provides the basis for a reasoned account of human duties is spectacularly ambitious; even if it fails in some ways it is worth the closest attention. In this chapter I aim to give as coherent an account of that attempt as I can offer, although I shall say nothing about the connections Kant draws between practical and theoretical reason (see Neiman 1994 ; O'Neill 1989 , chap. 1,

and 1992; Guyer 2000 , chap. 2). Since practical reasoning aims to shape and select action, I begin with a short account of Kant's views on action.

1. Practical Reasoning and the Agent Perspective

Agents use practical reasoning to shape or guide their future action. Since practical reason has to bear on action yet to be done, it cannot bear on *act tokens:* there are no relevant, individuable act tokens at the time that practical reasoning takes place. So practical reasoning has to bear on *act types* (including *types of attitude*). It might be used to provide reasons for thinking that certain types of action or attitude are required or forbidden, recommended or inadvisable.

As Kant sees it, types of action are specified by act descriptions, while normative claims are expressed in principles that incorporate act descriptions. Agents may consider, explore, test, adopt, or reject practical principles. Kant speaks of the principle an agent adopts as *determining an agent's will:* it fixes—in the sense of *making determinate* rather than of *causing*—some aspects of the action or attitude to which an agent is committed. Kant calls the more significant determinations of an agent's will "maxims" (G 4:402n, 4:421n; CPrR 5:19). Kant sees maxims as the practical analogues of *beliefs*. Individuals may *believe* a theoretical claim or proposition at or for some time; they may make a practical proposition their *maxim* at or for some time. Like beliefs, maxims have

propositional structure and content, so are apt for reasoning. Kant's most basic thought about practical reason is that reasoning can bear on action because it is formed or shaped by maxims, which have propositional structure and content.

In classifying only the more general principles that agents adopt as maxims, Kant is true to the etymology of the term. A maxim is the *maxima propositio*, a high- or highest-level proposition determining an agent's will at some time. The maxim an agent adopts will govern and inform other more specific decisions and aspects of his or her action or attitudes. For example, anybody who has adopted a maxim of not deceiving others is likely to express it in refraining from lying, in restraint from gossip, in care about checking facts, and many other ways. Maxims can be for the long or the short term. They may be deeply entrenched in an agent's character or in the constitution of a collective agent (R 6:89), or adopted in face of a particular situation or for a short period. Kant usually speaks of agents as adopting a range of distinct maxims at a given time, but in a few passages he suggests that we can speak of the *deepest* or *most fundamental* principle of a

person's character as a single maxim that governs their adoption of other more specific maxims, and thereby their entire life, often for a prolonged period. [2] However, maxims are always adopted and discardable, so they are something for which agents are responsible and which they might change.

There have been many discussions of the ways by which we may know Kantian maxims (Herman 1993 ; O'Neill 1989 , 1996 ; Timmerman 2000). On some readings of Kant, agents are conscious of their maxims and know them by introspection. However, this interpretation is hard to reconcile with Kant's views of the limits of human self-knowledge: he claims that we are not transparent to ourselves, but rather opaque (G 4:406–12). Hence the notorious passages in which he points out that we cannot *know* whether there has ever been a *truly* loyal friend, or that we have acted *purely* for the sake of duty (G 4:407–8; Baron 1995). On other views, maxims are ascribed to agents on the basis of a range of evidence, of which introspective evidence forms at most part. Both introspective and ascriptive views of knowledge of maxims focus on the *retrospective*, usually *third person* task of identifying the maxim(s) on which an act was done. In concentrating on the methods by which we can discover which maxim(s) are held by an agent at some time, they overlook the fact that the main task of practical reasoning is *prospective*. In reasoning about action, we consider maxims that could be adopted, viewing them as principles or prescriptions for whose adoption reasons might—or might not—be found. A prospective and prescriptive account of maxims provides the basis not for discovering which maxim(s) an agent actually adopts on a given occasion, but for determining which maxim an agent has reason to adopt. The basic task of practical reasoning is to guide action rather than to adjudicate past acts.

Since Kant takes a prospective and practical approach to reasoning about action, he can largely avoid the problem of showing how we are to discover agents' maxims or work out the “relevant” description for any act. When we try to assess action retrospectively, we have to work out which of many descriptions and principles satisfied by a given act is relevant for assessment. If assessment is undertaken for some specific purpose, such as

financial audit or legal judgment, conformity or lack of conformity to relevant descriptions can be judged. And if the main aim of moral assessment were to judge agents' maxims retrospectively, we would apparently need a general way of finding out what maxims they have "really" adopted. Kant says a good deal about retrospective judgment of action in his discussions of reflective judgment, [3] but (unlike some leading contemporary writers on moral perception, appraisal, or judgment [Wiggins 1987 ; McDowell 1996]) he assigns priority to prospective, practical reasoning rather than to a retrospective, spectator perspective on action or ethics.

The potential maxims on which agents may bring practical reasoning to bear may be of many sorts. They do not have to be morally admirable: for example, Kant discusses a group of hard-bitten "sophistical" maxims of political expedi

end p.95

ency. [4] A neutral view of the moral status of maxims is appropriate: if practical reasoning is to show why we should adopt some rather than other principles as maxims, setting prior limits on which principles are to be adopted as maxims would beg questions. Nor does Kant propose a method for ensuring that agents assemble possible maxims for consideration. Agents may well fail to consider some principles that they have reason to adopt as maxims. However, Kant takes it that agents are not likely to be systematically blind to the central principles of duty, which are repeatedly relevant to decisions and action.

2. Hypothetical Imperatives

Some commentators have imagined that since Kant holds that practical reasoning can set unconditional requirements (so is a *categorical imperative*), he must deny that practical reasoning sets conditional requirements. Such a position would, of course, be absurd. Reasoning cannot guide action—cannot be *practical*—without taking a view of the connections between types of action and types of effect, between means and ends, between action and world.

Kant's account of instrumental reasoning is adjusted to his view that reasoning bears in the first instance on potential determinations of the will or maxims, and thence on action. He speaks of the fundamental principle of instrumental reasoning as the *principle of hypothetical imperatives*, and formulates it as an abstract principle for rational willing. Hypothetical imperatives "represent the practical necessity of a possible action as a means to achieving something else that one wills (or that is at least possible for one to will)" (G 4:414; also G 4:414–19; CPrR 5:19–20; Hill 1992 , chaps. 1 and 7; Wood 1999 , chap. 2). This principle requires commitment to the maxim "Whoever wills the end also wills (insofar as reason has decisive influence on this actions) the indispensable necessary means to it that are within his power" (G 4:417).

Kant argues that those hypothetical imperatives that supposedly guide the pursuit of happiness do not, strictly speaking, set *requirements* for action. Happiness is an indeterminate ideal—an "ideal of the imagination" (G 4:418)—so means-end reasoning

directed at the pursuit of happiness yields at most approximate, *pragmatic imperatives* or *counsels of prudence*. These commend ways of living that generally conduce to happiness, such as "frugality, courtesy, reserve, and so forth" (G 4:418). Other hypothetical imperatives are grounded in technical and causal requirements. They set genuine requirements, but only for those pursuing specific ends. Kant speaks of them as *technical imperatives* or *rules of skill*.

Rules of skill are morally neutral: knowing the effects of poisons is as useful to physicians as it is to poisoners (G 4:415).

Kant's account of means-ends reasoning is clearly less ambitious than those favored in many contemporary accounts of rational choice. He does not assume that we can list "options" exhaustively, that we have complete or even very extensive knowledge of causal connections or probabilities, or that there is a metric for ranking or aggregating the value of ends. His account of instrumental reasoning does not provide enough structure for much in the way of judgments of efficiency, and says little about the distinction between willing necessary and sufficient means. He also does not place the entire burden of practical reasoning on means-ends reasoning.

3. Categorical Imperatives: Universal Law

In every area of life, instrumental reasoning is undertaken in pursuit of agents' chosen ends: these choices orient and require determinate means-ends reasoning. However, on Kant's account, specifically moral reasoning needs more than a combination of chosen ends and means-ends rationality. Those who hope to get by on this basis will defend some form of "heteronomy in ethics." [5] They choose to make some intrinsically arbitrary ends the basis of their ethical reasoning (see also Hooker and Streumer, chap. 4, Smith, chap. 5, and McNaughton and Rawling, chap. 7, this volume). Their choices may variously endorse or defer to self-interest, religious dogma, established ideology, community "values," or some version of self-development. Even those who propose an account of human flourishing or happiness as the foundation and context for moral reasoning do not escape this arbitrariness, given that they lack an adequately determinate account of human happiness or flourishing. Kant's claims about the arbitrariness of heteronomous ethics are not hard to appreciate. Yet the thought that there could be any way of avoiding heteronomy in ethics looks quite implausible. How can there be *any* unconditional requirements on action? What reason have we to think that there is anything that could count as the rejection of heteronomy in ethics—as *autonomy in ethics*?

Kant's answers to these questions can be found mainly in the *Critique of Practical Reason* and the *Groundwork of the Metaphysics of Morals*, although significant additional analyses and comments occur throughout his later writings on ethics, politics, religion, and history. His thoughts are often difficult to follow, in

part because he offers several seemingly distinct (yet supposedly equivalent) formulations of the supreme principle of practical reasoning, the categorical imperative. I shall begin with some comments on the well-known formula of universal law (FUL), the formulation most closely linked to Kant's attempted vindication of practical reason. Finally, I will comment on some of the other formulations and on Kant's claims about their equivalence.

FUL proposes a doubly modal requirement as the basis for moral reasoning. Its best-known version runs: *Act only in accordance with that maxim through which you can at the same time will that it become a universal law* (G 4:421). Many commentators—not least among them John Stuart Mill (1962 , 254)—have claimed that this is too little. Despite his protestations, Kant must in the end go beyond the modal demand set out in FUL if his conception of practical reason is to guide action. Surely almost any maxim that *can* be adopted by an individual agent *can* also be adopted by all agents? The only maxims that cannot be adopted by all are, it may seem, those that are intrinsically incoherent, and so cannot really be adopted by individuals either (e.g., a maxim of being a popular recluse). Even maxims that refer to positional goods (e.g., a maxim of becoming richer than everyone else) could be *adopted* by all, although failure to achieve the maxim's aim would be guaranteed for all but the most successful agent. It seems that a requirement to act only on maxims that can be willed as universal laws cuts little ice.

Such criticisms oversimplify Kant's account of FUL. Kant views FUL as demanding that an agent be *able to will* the maxim he or she proposes to adopt "as a universal law." Willing is not merely a matter of *thinking* or of *wishing* some practical proposition: it is matter of making a certain proposition one's maxim, of adopting it as a determination of one's will, and this means engaging with the demands of the principle of hypothetical imperatives (G 4:394). Willing a maxim "as a universal law," although only a hypothetical test (I cannot *literally* will for others), requires agents to consider whether everyone *could make the proposed maxim the determination of his or her will*, so must engage (hypothetically!) with the demands of the principle of hypothetical imperatives. When I will a maxim "as a universal law" I do not (even hypothetically!) will that *everyone act on the maxim* at some moment, or at all moments. Very many practical principles, including principles of great and of trivial moral significance, cannot be acted on by everybody at any one time and place. If everyone tries to help a drowning child or to swim across a river simultaneously, overcrowding and mutual obstruction would guarantee that some cannot act. Kant's question is whether *everyone could will a maxim*. There is nothing incoherent in everyone *willing* to rescue a drowning child or to swim across a river, or even a particular river—but not everyone can act simultaneously on either principle in a given location. (Anyone who wills a child's rescue will remain on the shore if wading in would obstruct the rescue.)

At this juncture it may seem that the demand that agents act only on maxims

that they can will as universal laws fails to guide action for quite a different reason. Its problem is not that it is "too specific," that it rules out each determinate act as impermissible on the grounds that nothing can be done by everyone at a given time and place. Its problem is that it is "too general," since it requires only that agents can *adopt*

maxims that could be adopted by all, so perhaps rules out nothing. Surely, if *anyone* can adopt a practical principle, then *everyone* can adopt it.
Kant thinks that this is not the case. Some principles can readily be adopted by a given agent or a minority of agents, *but only on the assumption that they are not adopted by all others*. For example, adopting a maxim of promising falsely commits an agent to supporting means to promising falsely, hence to maintaining enough public trust for promises to gain acceptance. But willing false promising "as a universal law" (*per impossibile*) commits an agent to willing the consequences of universal false promising, which include the destruction of trust, hence is incompatible with willing any reliable means to false promising—for oneself or for others. That is why Kant thinks that we *cannot* will false promising, coercion, and many other types of action that victimize others "as universal laws." He points out that we do not even pretend to do this: "If we now attend to ourselves whenever we transgress a duty we find that we do not in fact will that our maxim should become a universal law—since this is impossible for us—but rather that its opposite should remain a law universally: we only take the liberty of making an exception to it for ourselves or even just for this once" (G 4:424).
But why is rejecting maxims that cannot be willed "as universal laws" a formula for identifying *duties?* Even if we accept that FUL appears to show that some maxims of false promising cannot be "willed as universal laws"—cannot be universalized—is this more than a curiosity? Why should we think that FUL picks out principles of action that we ought not to adopt? And even if it does, why should we think that it is a version of the "supreme principle of morality"? And why, above all, should we think that this curious formula is the fundamental principle of practical reason? Further, why should we think that the other formulations of the categorical imperative, which Kant claims are equivalent to FUL, are also versions of the supreme principle of morality and of the fundamental principle of practical reason?

4. Universal Law as a Principle of Practical Reason

Before turning to the other formulations of the categorical imperative I shall consider why Kant sees FUL as (one version of) the supreme principle of practical
end p.99

reason. Kant's claims about practical reason can seem bombastic: phrases such as "the supreme principle of practical reason" arouse suspicion. Yet Kant is very cautious about what reason, including the practical use of reason, can provide.
This caution about the authority of reason is expressed in vivid terms in the prefaces and introduction of the *Critique of Pure Reason*. As Kant sees it, we do not even know where or how to begin the "tasks of reason." Reason is not *given* to us; it is not "whole and complete in each of us," as Descartes supposed (Descartes 1984 , 112). We constantly find ourselves using ways of thinking and acting that we speak of as reasoned; but we also find that daily reasoning goes horribly wrong. Particularly when we seek to extend our reasoning beyond experience and aspire to reach metaphysical conclusions, we

constantly find that “we have to retrace our path countless times, because[reason] does not lead where we want to go, and it is so far from reaching unanimity in the assertions of its adherents that it is rather a battlefield.Still more, how little cause have we to place trust in our reason if in one of the most important parts of our desire for knowledge it does not merely forsake us but even entices us with delusions and in the end betrays us!” (CPR B:xv; see also A:viii; B:xiv). As is well known, Kant's central proposal for bringing the weary battles of metaphysics to an end is to insist that human reason cannot give us a route to knowledge that reaches beyond possible experience and the presuppositions of experience.

He insists that we must be wary of what we take to be the powers of human reason and put them too to the test. Reason itself must be judged and scrutinized: “Reason should take on anew the most difficult of all its tasks, namely that of self-knowledge, and institute a court of justice by which reason may secure its rightful claims, while dismissing all its groundless pretensionsthis court is none other that the critique of pure reason itself” (CPR Axii). Yet the idea of vindicating reason by appeal to a “court” or “tribunal” can seem absurd. What procedures of judging could have the status to determine what does and what does not qualify and count as reasoning? Since there is no more general or fundamental claim to authority in organizing thinking and acting than an appeal to reason, how can reason itself be judged? How can any “tribunal of reason” have standing to judge the competence and limits of reason? And how, if reason cannot be judged, can appeals to reason gain any authority? Perhaps the daring and demanding thought that reason lacks credentials leads to the conclusion with which postmodernists flirt: perhaps what passes for reason among us has no authority at all (O'Neill 1989 , chaps. 1 and 2).

Kant's move at this point [6] has become more familiar during the last thirty years through the work of John Rawls, Jürgen Habermas, Thomas Scanlon, and others. He proposes a *justification* or *vindication* of reason, rather than a *proof* or *foundation* for reason. Justification differs from proof in that it is directed to some

audience; and unconditional justification must be directed to audiences without assuming that they meet any specific conditions, so must be directed to all agents. Procedures that can serve to reach a limited audience who share a particular conception of the world (perhaps embodied in shared beliefs or prejudices, in shared community or citizenship) can at best have restricted authority. [7] At most they can provide a basis for parochial, conditional reasoning. By contrast, unconditional reasons must be fit to reach “the world” (WE 8:38; in other translations, “the world at large”), rather than a restricted audience with whom an agent happens to share much. Reasoning that can reach only a restricted audience is incomplete or conditional: Kant calls it *private* or *heteronomous* reasoning. [8] Reasoning that can reach the world at large is unconditional: Kant calls it *public* or *autonomous* reasoning. The fundamental move on which Kant's vindication of reason depends is the requirement that it be *fit for universal use*, rather than adapted to some restricted audience. By contrast, appeals to any local, restricted consensus or agreement would provide only parochial, limited, and conditional reasons for action.

Those who aspire to offer reasons to unrestricted audiences, and so assume no prior conditions that secure agreement, face a hard task. Their attempts at reasoning will fail unless they ensure that their proposals are accessible to an unrestricted audience. [9] Those who propose *reasons to accept certain beliefs* to "the world at large" must ensure that all others can in principle *follow* the moves that they make in presenting their thoughts: they must aim for intelligibility, without overtly or covertly assuming prior agreement. Those who propose *reasons for acting* to "the world at large" must aim not only for intelligibility: they must propose principles of action that others not merely can follow in thought, but could adopt as principles of action. I do not offer reasons for action to all if I propose principles of action that I know some others *cannot* adopt. Another way of putting this requirement is to say that those who wish to offer practical reasons to an unrestricted audience must *act only in accordance with that maxim through which [they] can at the same time will that it become a universal law* (G 4:421). FUL states (a version of) the supreme principle of practical reason because it states the condition for anything to count as a reason for action for an audience about whom we make no special, restrictive assumptions. It is therefore a requirement for giving unconditional reasons for action, a categorical imperative for the adoption of maxims. The underlying reason why Kant thinks that practical reasoning has to propose maxims that are fit to be universal laws is *that no other maxims can coherently be offered as reasons to all*. At most they could coherently be offered as reasons to a restricted range of agents who accept *some further, rationally ungrounded assumption or attitude*.

5. Universal Law and Moral Duties

If practical reason amounts to a demand to act only on principles that have the form of law, that can be principles for all, it offers only an *indirect* and *incomplete* standard for morality. I shall consider the implications of offering an *indirect* standard of morality in this section, and those of offering an *incomplete* standard in section 6.

FUL provides only an *indirect* standard for guiding action since it identifies principles that ought to be rejected, rather than principles that ought to be adopted, or that it would be good to adopt. However, knowing that some principles ought to be rejected can guide action. If I know that a principle of revenge cannot be universalized, I have reason to reject a maxim of revenge. If I know that a principle of coercing others cannot be universalized, I have reason to reject a maxim of coercion.

Such reasons are, however, less than conclusive. It is sometimes impossible, hence not obligatory, to refrain wholly from types of action whose maxims we have reason to reject. For example, Kant, as is well known, doubts whether human society can exist without some coercion: he is neither a pacifist nor an anarchist. Rather he argues that the very principle of rejecting coercion cannot be respected without using some coercion. A just political system not merely *may* but *must* coerce to limit coercion, and more generally hinder freedom in order to limit hindrances to freedom (MM 6:230–33; Mulholland 1990 ; Flikschuh 2000 ; Timmons 2002 ; Guyer 2002). In acting on a maxim of rejecting coercion we have to deploy certain very specific forms of coercion. In identifying principles that we cannot universalize, so have reason to reject, Kant does not

commit himself to principles of duty that are blind to circumstances and realities or deny the possibility of conflicts between the various claims of (one or more) moral principles (Herman 1993 ; Baron 1995 , chap. 3; O'Neill 2002a).
FUL can be used to identify further principles of action that cannot be willed as universal laws. Principles of doing violence, of victimizing, or of undermining others' capacities to act in other ways, cannot coherently be willed as universal laws because their universal adoption (*per impossibile*) would predictably undercut the possibility of adopting those very principles for at least some others (Herman 1993 ; Baron 1995 ; O'Neill 1989). Those who adopt such principles in effect view themselves as enjoying exceptional moral status: they may know that their maxims cannot serve as principles for all, but see this as irrelevant because they do not view all others as their moral equals. On Kant's view, reasons for rejecting principles of action that cannot be universalized enable us to identify the fundamental principles of duty.
Kant divides basic duties into two classes, identified respectively by what he

terms the *contradiction in conception* and *contradiction in the will* applications of FUL:
We must *be able to will* that a maxim of our action becomes a universal law: this is the canon of moral appraisal of action in general. Some actions are so constituted that their maxims cannot even be *thought* without contradiction as a universal law of nature.In the case of others that inner impossibility is indeed not to be found, but it is still impossible to *will* that their maxim be raised to the universality of a law of nature because such a will would contradict itself. (G 4:424)
The *contradiction in thought* (or *in conception*) test identifies maxims of *strict (narrow, perfect) duty*, including duties of justice. It picks out maxims of action that cannot coherently be thought of as principles for all. The *contradiction in the will* test identifies maxims of *wide (imperfect) duty*, including duties of virtue. It picks out maxims that can coherently be *thought of* as principles for all, but cannot be *willed as* principles for all in a world of interacting agents. For example, Kant thinks that *taken in isolation* a maxim of mutual indifference or a maxim of neglecting to develop any skills or talents could consistently be universalized: they pass the *contradiction in conception* test. But nobody can consistently will that these principles be universally adopted in any world of interacting agents whose members must (by the fact that they are instrumentally rational) will to receive others' help and to rely on others' skills if and when their own are insufficient: they fail the *contradiction in the will* test (Herman 1993 , chaps. 3, 6, and 7; Baron 1995).
The contradiction in the will test can be looked at in more than one way. On a minimal reading it is a matter of prudence. No reasonable agent who acknowledges *her own* finitude and vulnerability, and consequent inability to achieve all her ends unaided, can coherently will to be part of a world of agents who are indifferent to others' needs or who systematically neglect to develop human skills: in doing so she would flout the demands of instrumental rationality. On a wider reading, no reasonable agent who acknowledges *others'* finitude and vulnerability and so knows that nobody can achieve all their ends without help can will universal indifference to human needs or to the development of human skills. Willing universal mutual indifference amounts to willing a world in which

agency and capacities fail for some or many, so undermining action. For similar reasons, no rational agent can consistently will universal failure to develop skills. Willing universal failure to develop skills amounts to willing a world in which some or many find their capacities to act at risk. Kant, of course, acknowledges that some *individuals* may get away with large amounts of indifference to others, and with failure to develop skills: free riders often get away with it. He denies, plausibly enough, that we can will either sort of free riding as a universal law for a world of interacting agents.
end p.103

6. Universal Principles and Judging Cases

Arguments from FUL to these broad principles of duty also offer a very *incomplete* standard for morality. The perennial allegation that Kantian practical reason is *too abstract* or *too formalistic* sees this incompleteness is a serious defect (Mill 1962 , chap. 1). The charge of *formalism* is that Kant identifies only very general principles of duty, whereas we need to know just what to do in particular circumstances. Kant himself pointed this out: "A physician, a judge or a ruler may have at command many excellent pathological, legal or political rules, even to the degree that he may become a profound teacher of them, and yet, none the less, may easily stumble in their application. For, although admirable in understanding, he may be wanting in natural power of judgment. He may comprehend the universal *in abstracto* and yet not be able to distinguish whether a case *in concreto* comes under it" (CPR A:134; B:173).
Discussions of judgment, including practical judgment, are ubiquitous in Kant's writings. He never assumes agents can move from principles of duty, or from other principles of action, to selecting a highly specific act in particular circumstances without any process of judgment. He is as firm as any devotee of Aristotelian *phronesis* in maintaining that principles of action are not algorithms and do not entail their own applications. There has been a good deal of recent discussion of the details of Kant's views on practical, including ethical, judgment. (Herman 1993 ; Engstrom and Whiting 1996 ; O'Neill 2002a) His discussions of these topics are numerous, complex, and perhaps most abundant in *The Metaphysics of Morals*, which addresses many aspects of practical judgment, including deliberation, casuistry, and conflicts of obligation.
A second version of the charge that Kant's ethics is seriously incomplete objects that it is possible to devise artfully tailored maxims that can be willed as universal laws, but are morally obnoxious (Wood 1999 , chap. 7; Herman 1993). For example, instead of testing a general maxim of revenge or false promising, as Kant does, we might test maxims permitting persons of specific sorts or status to exact revenge or to promise falsely, knowing that deception or revenge by narrowly specified categories of agents could be willed as universal laws without contradiction. Such proposals for undercutting the implications of Kant's ethics overlook two difficulties. First, they are based on according some people exceptional moral status, denied to others, so they reject Kant's fundamental view that human beings are moral equals. Second, those who advance them fail to note that basic principles of duty do not fall away when we consider a more closely specified

line of action. The general duty to reject revenge does not fall away just because a proposed act of revenge falls under more specific descriptions and principles; a
end p.104

general duty of fidelity does not evaporate because an agent is tempted by a scam that is open to few.

7. Universal Laws and Ends in Themselves

Perhaps the most mysterious feature of the categorical imperative is that Kant formulates it in several distinct ways that look quite different, but which he claims are equivalent (G 4:436; Hill 1991 , 1992 ; Korsgaard 1996a , chaps. 2–4; O'Neill 1989 , chap. 7; Wood 1999 , esp. chap. 4). It is common to group the various formulations under four or five headings. Here I shall discuss the well-known formula of the end in itself (FEI) and formula of autonomy (FA), but say nothing specifically about the formula of the law of nature (FLN) and the formula of the kingdom of ends (FKE). There are two reasons, considerations of space apart. Most immediately, I bracket FLN because it is in many respects similar to FUL, and FKE because it is readily understood if sense can be made of FEI. Second, and more importantly, FEI and FA are the origins of the resonant contemporary moral ideals of *respect for persons* and of *autonomy*. Yet it is far from clear that either FEI or FA is equivalent to FUL, or that either is a version of the supreme principle of morality, rather than one moral principle among others. Still less is it obvious why either should count as a version of the supreme principle of practical reason.
FEI is formulated in the *Groundwork of the Metaphysic of Morals* in the words *So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means* (G 4:429). Kant himself views FEI as the most intuitive version of the categorical imperative; many recent accounts of Kantian ethics go further and dismiss FUL in order to concentrate on FEI's demands for respect for persons (G 4:436; Hill 1992 ; Wood 1999 , chap. 4; for criticism, see Regan 2002). Certainly, FEI does not look equivalent to, or even like, FUL. Yet if no equivalent reading can be found, Kant's ethical theory will fall apart: the deepest arguments used to justify the categorical imperative are directed mostly at FUL, so can provide grounds for FEI only if it can be read as equivalent to FUL (O'Neill, 1989 , chap. 7).
The fundamental difference between FUL and FEI is that FUL constrains what agents should (may, may not) do, whereas FEI constrains how agents should (may, may not) be treated. The two formulas consider requirements on action respectively from the agent's and the recipient's point of view. FEI, however, is more explicit than FUL about those on the receiving end of action. It requires right
end p.105

treatment *of humanity*, [10] whereas FUL requires action on maxims that can be principles *for all*. If FUL and FEI are to be read as equivalent, it is essential to read them using a

common account of the scope of ethical concern. For example, it is common to read FUL as requiring action on maxims that can be principles *for all human agents*, so using the same view of the scope of ethical concern as is explicit in FEI. Alternatively, both FUL and FEI could be read as setting requirements on action by and treatment of rational natures, human or other (Hill 1992 , chap. 2).

Even when FUL and FEI are read using a common view of the scope of ethical concern, their equivalence is not obvious. Why should acting on principles that can be principles for all amount to treating others as ends and not as mere means? The issue can best be approached by considering first what is meant by "treating others as mere means." Kant distinguishes interaction with others that respects and preserves their capacities as agents, in which persons permissibly use one another as *a means*, from action that uses unwilling others as *a mere means*. We use others as mere means when we treat them not as agents but as things or tools, as something to be "used by this or that will at its discretion" (G 4:428). Those who adopt such maxims do not always damage others' agency (they may lack opportunity or power), but the *standard* results of such action damage agency. We cannot will any of numerous maxims of victimizing as universal laws: in doing so we would will results that would undercut the means for like action for some or for many. Those who treat others as *mere means* act on maxims that they *cannot* will as universal laws. This argument does not establish the equivalence of the FUL and FEI, but only that the first part of FEI—*do not use [others] merely as means*—follows from FUL. Strict duties can be derived either by applying the *contradiction in conception* test aspect of FUL, or by adhering to the part of FEI that prohibits treating others as mere means.

The second part of FEI enjoins treating others as "ends in themselves" and corresponds to the contradiction in the will test application of FUL, by which maxims are tested not for inner coherence, but for fitness as laws for a possible world of agents. In treating others as ends in themselves, we treat them as persons, as beings who have objective worth: this requires more than refraining from treating them merely as means. We treat others as ends by acting in ways in which a world of agents can be sustained: by acting on maxims that can form part of a system of maxims that can be willed without contradiction, that harmonize with the necessary conditions for sustaining a world of agents (G 4:430–31).

Provided that FUL and FEI are read using a common view of the scope of ethical concern, the difference between them is in the end a matter of perspective. FUL is a test for *ways of acting* that could be adopted by all agents in a world of agents; FEI is a test for *ways of being treated* that could be accepted by all agents within some world of agents. Within a world of agents, the perspectives of agency and recipience consider the same set of possible actions from different points of view. Both formulas state procedures for rejecting maxims whose universal adoption would undermine at least some others' possibilities of like action.

8. Universal Laws and Autonomy

In many ways the formula of autonomy (FA) is closer to FUL than it is to FEI. In the *Groundwork* Kant writes, "The principle of autonomy isto choose only in such a way that the maxims of your choice are also included as universal law in the same volition" (G

4:440). [11] The puzzle is not so much that Kant sees FA as a version of the categorical imperative—it is, after all, quite close to FUL (but see Wood 1999 , chap. 5)—but rather to understand why he speaks of it as the formula or principle *of autonomy*. How can he view autonomy as basic to morality, or claim that "morality is thus the relation of actions to the autonomy of the will" (G 4:439)? [12]

This puzzle can be resolved by distinguishing Kant's conception of autonomy adequately from contemporary conceptions. Most current discussions of autonomy see it as a property of individual agents, which they may have to a greater or lesser degree and may express in some domains of life more than in others. Contemporary conceptions of autonomy generally equate it with forms of personal independence or self-expression, or with acting on certain distinctive, supposedly "autonomous," preferences (O'Neill 2002b , chaps. 2 and 4). Such conceptions of individual autonomy are widely discussed, and their ethical merits widely disputed. These discussions are irrelevant to understanding Kantian autonomy. Kant never writes of autonomous *selves* or *persons* or *individuals*. He predicates autonomy *of reason,of ethics,of principles,of willing* (Hill 1991 , chap. 4). Although contemporary advocates of individual autonomy often claim Kantian ancestry for their ideas, the claim is bogus, made plausible only by distorting Kant's conception of autonomy in major ways.

For Kant the idea of autonomy captures the two central aspects of his account of practical reason and of ethics: that duty is a matter of acting on principles or laws, and that those principles or laws should not be derived from arbitrary starting points. Principles are autonomous on Kant's view not because they express some particularly striking or independent personal decision or attitude, but because they are *not derived from elsewhere*. "Self-legislation," as Kant writes of it, does not mean that each self or each agent chooses or "legislates" moral principles for all (coherent only where there is some extraneous, hence unreasoned, source

of coordination). He understands self-legislation as *the law giving of reason* rather than as *the law giving of individuals*. He remarks that "it is requisite to reason's lawgiving that it should need to presuppose only *itself*, because a rule is objectively and universally valid only when it holds without the contingent, subjective conditions that distinguish one rational being from another," and he states that autonomy requires a "*law-giving of its own* on the part of pure and, as such, practical reason [which] is freedom in the *positive* sense" (CPrR 5:21, 33). For Kant the element *self* in the term *self-legislation* indicates a *reflexive* claim that lawgiving, and hence principles and maxims, not be derived "from elsewhere"—that is, from sundry arbitrary assumptions or conditions. Kantian autonomy is a matter not of self-expression by individual selves who "legislate," but of agents choosing maxims—"laws"—that are nonderivative, so nonconditional.

This conception of autonomy as willing that does not appeal to arbitrary (bogus, or at best conditional) "authorities" is the basis of Kant's contrast between *autonomous ethics* and positions that advocate forms of *heteronomous ethics*. The proponents of heteronomy in ethics derive supposed moral principles from "authorities" such as Church or state, ideology or market forces, public opinion or personal preference, individual choice or majority vote. The practical reasoning deployed in heteronomous ethics is therefore

essentially instrumental or conditional: it is simply a matter of choosing action that implements the standards of the supposed "authority" effectively and efficiently.
As Kant sees it, "If the will seeks the law that is to determine it *anywhere else* than in the fitness of its maxims for its own giving of universal law—consequently if, in going beyond itself, it seeks this law in a property of any of its objects—*heteronomy* always results" (G 4:441). If we are to offer reasons for action that are relevant to *all* others, we must begin by ensuring that we offer reasons they *can* offer and consider, accept or refuse. We cannot do this by relativizing our reasons for action to any "authority" that some—or many—may have no reason to accept and may in fact reject. Rather than trying to build personal independence or other conceptions of individual autonomy to ethics, Kant argues that we should ensure that what we propose to others is based on principles that are *fit* to be laws, hence on principles that they at least *could* adopt. This seemingly slender modal requirement demands that we reject all forms of heteronomous ethics in favor of acting on principles that are fit to be universal laws, hence also that we not treat others as mere means or as less than ends. Kant sums up these views on reason and morality in the striking claim that "the moral law expresses nothing other than the *autonomy* of pure practical reason" (CprR, theorem IV, 5:33).
end p.108

NOTES

1. Kant citations use the abbreviations indicated in the bibliography and the page numbers of the Prussian Academy of Sciences edition, which are given in the margins of each translation.
2. See the discussion of making either the moral law or self love one's most fundamental maxim and so determining one's basic disposition, R 6:22ff.
3. In reflective judging "only the particular is given, for which the universal is to be found" (CJ 5:180).
4. They include "Fac et excusa," "Si fecisti, nega," and "Divide et impera," PP 8:374–75.
5. See the discussion on heteronomy and autonomy in ethics at G 4:440–44 and that of public and private uses of reason in WE 8:35–42.
6. Kant makes the same move in his account of freedom, for which he offers a *vindication* or *defense* but no *proof*. See G 4:445–63.
7. Kant would therefore be unconvinced by John Rawls's conception of the reasonable as a form of public reason shared by fellow citizens within a bounded democratic society. He would think it inadequate as vindication of reason because it presupposes and does not justify bounded territories, citizenship, and democracy.
8. See G 4:440–44 and WE 8:35–42.
9. Accessibility is not the same as motivational sufficiency. For discussion of the motivational claims of Kant's view of practical reason, see Korsgaard 1996a , chap. 11.
10. Kant discusses the claims of *humanity*, of *rational nature*, and of *sentient nature* in many passages, especially in G 4:448 onward and in MM. For discussion of his humanism, and of charges of speciesism leveled against him, see Wood and O'Neill 1998 .

11. See Wood 1999 , 163–64, for a listing of versions of FA and an argument that FA is quite distant from FUL.
12. Even more strikingly, he equates it with practical reason: the power to judge autonomously—that is, freely (according to principles of thought in general)—is called "reason" CF 7:27.
end p.109

Chapter 7
DUTY, RATIONALITY, AND PRACTICAL REASONS

David McNaughton
Piers Rawling

Some authors see tight connections among duty, rationality, and practical reasons (see, e.g., O'Neill, chap. 6, this volume). We shall present a view on which the connections are relatively weak. The use of these terms is partly stipulative (see, e.g., Scanlon 1998 , chap. 1), but ours conforms to some aspects of common usage, and it serves to keep track of crucial distinctions. When Davidson (1980 , chap. 1) speaks of reasons, he is referring to psychological states that can be cited in explaining an action. Smith (1994 , 96) refers to these states as "motivating reasons." Practical reasons, as we shall think of them, however, are not psychological states. They are facts, such as the fact that the rubbish bin is full. This is a non-normative fact, but, to use Parfit's phrase (1997 , 124), it has "normative significance": it is a reason for you to do something, namely take the rubbish out. Rationality we see as a matter of consistency. Failing to notice that the rubbish bin is full need not be a rational failure. Duty is a matter neither purely of rationality nor of practical reason. On the one hand, the rational sociopath is immoral. But, on the other, morality does not require that we always act on the weightiest moral reasons—we may not be reasonably expected to know what these are.
end p.110

We have mentioned the normative force of practical reasons, but what of their motivational force? Williams (1981 , 1995a , 1995b) ties practical reason to both rationality and motivation (see also Hooker and Streumer, chap. 4, and Smith, chap. 5, this volume). Section 1 outlines his internalism; section 2 distinguishes other uses of the term. We discuss responses to Williams in section 3, and in section 4 we tentatively propose a view of duty that is neither purely subjective in Prichard's (1932) sense nor purely objective.

1. Williams's Internalism

Perhaps practical reasons, though not psychological states, are simply facts about them. Perhaps you have reason to Φ if and only if you desire to, because a reason to Φ just is

the fact that Φ-ing is desired. On such a view, reason-talk may not be devoid of normative force (one might be criticizable for, say, not doing what one desires most), but it is too limited for Williams's taste (1995a , 36). On his view, an agent has an "internal" reason to Φ only if she would arrive at a motivation (or desire—we use the terms interchangeably) to Φ were she to deliberate rationally from her current "motivational set," *S*, where the latter has been corrected to eliminate false beliefs and include all relevant true beliefs (Williams 1981 , 102–3; 1995a , 36; see also Smith 1994 , 156, and Parfit 1997 , 100). (*S* includes not only ordinary desires, but also "such things as dispositions of evaluation, patterns of emotional reaction, personal loyalties, and various projectsembodying commitments of the agent" [1981, 105]; and reasons are not "all-in" [1981, 104]: one can have a reason to Φ while having stronger reasons not to.)
What is it to deliberate rationally? Parfit (1997 , 101) dubs the two relevant notions "procedural" and "substantive" rationality (see Hooker and Streumer, chap. 4, this volume). To be substantively rational, one must have certain concerns—such as prudential, and, perhaps, moral concerns. Procedural rationality builds in no such requirements, and is the defining notion for internal reasons. Williams claims that all reasons are internal: that is, if *A* would not arrive at a desire to Φ were she to deliberate in procedurally rational fashion from her current *S*, relevantly corrected, she has no reason to Φ. Furthermore, and crucially, if *A* has a reason to Φ, that is *because* she would arrive at a desire to Φ were she to deliberate in procedurally rational fashion from her current *S*, relevantly corrected (see Parfit 1997 , 102). Your reason to take the rubbish bin out is the fact that it is full. But this fact is a reason for you only because if you were apprised of it, you would, via (a presumably short) procedurally rational deliberation, acquire the desire to take it out.

The externalist claims, by contrast, that an agent can have a reason to, say, take the medicine, even if she has no internal reason to do so (Williams 1981 , 106; Parfit 1997 , 100–101). The reason is that the medicine will cure her. Furthermore, on the external account, even if she arrives at a desire to take the medicine via a rational procedure, she does not have a reason to take the medicine because of this (cf. Plato's *Euthyphro*). The fact that she has a reason to take the medicine stands metaphysically independent of her *S*, just as the fact that she has reason not to affirm the consequent stands metaphysically independent of her grasp of this fact, and her ability to desist. As on the internalist view, your reason to take the bin out on the externalist view is the fact that it is full. But this fact is a reason for you to take it out independently of whether, if you were apprised of it, you would, via rational deliberation, acquire a desire to do so.
Before pursuing this internalist-externalist debate, we pause to say something about other uses of the term "internalism." (For further discussion and references, see Brink 1989 , 37ff.; McNaughton 1988 , 21ff.; Parfit 1997 ; Smith 1994 , 60ff.)

2. Other Internalisms

We have so far been discussing internalism as a claim about the conditions a fact must satisfy in order to be a reason for an agent to act—roughly speaking, the claim is that a

fact is a reason for *A* to Φ only if, were the agent to believe that fact and various others, she could arrive at a motivation to Φ via a rational procedure. This is a form of reasons internalism (Parfit 1997 , 100). Williams sometimes speaks as if this reasons internalism is a meaning claim: "I think the sense of a statement of the form `A has a reason to Φ' is given by the internalist model" (1995a , 40). At the least, the idea is that there is a conceptual or "internal" connection between reasons and motivation.
Sometimes, however, "internalism" refers to an internal connection between psychological states. Consider Smith's (1994 , 148)

(C2) If an agent believes that she has a normative reason to Φ, then she rationally should desire to Φ.

(Propositions are listed in the appendix in their order of appearance.) Smith takes (C2) to be a conceptual truth and sees it as following from Williams-style internalism. Smith (1994 , 156) defines "full rationality" as requiring that the agent deliberate correctly, have no false beliefs, and have all relevant true beliefs (cf. Williams 1981 , 102–3). *A* has a normative reason to Φ only if *A*'s Φ-ing is desirable
end p.112

(1994, 95), which is analyzed (153) as: Φ-ing is what *A* would desire to do if she were fully rational. Thus Smith's version of Williams-style [1] reasons internalism is:

(W) *A* has normative reason to Φ in *C* only if *A* would desire to Φ in *C* if fully rational.

So, Smith claims (1994 , 177), if *A* believes that she has a normative reason to Φ, she believes, by (W), that she would desire to Φ if she were fully rational. And if *A* believes she would desire to Φ if she were fully rational, then she is irrational, by her "own lights," if she fails to desire to Φ.
There are difficulties here. Even if (W) is true, it does not follow that *A* knows this: [2] as Williams (1995b , 187–88) notes, *A* might *believe that* she has reason to Φ, while lacking the *belief that* she would desire to Φ if fully rational. Furthermore, an agent might *falsely* believe that she would desire to Φ if she were fully rational (cf. Williams 1981 , 103, [iii] [a]). Hence, despite the belief, she would be *rational* (in the "full" sense) in failing to desire to Φ.
It is important to keep distinct two separate requirements: that of truth in belief and that of internal consistency among one's propositional attitudes. Suppose an agent believes that she has reason to Φ while failing to desire to Φ. By (C2), this lack of desire is irrational. But "irrational" here surely refers to failure in internal practical consistency: suppose our agent *irrationally* believes that she has reason to Φ, then her motivational failure is irrational only in the sense that it fails to cohere with her belief. Furthermore, one can be fully rational (in Smith's sense) in failing to desire to Φ despite believing that

one has reason to Φ. If the listless agent has no desire to get out of bed while believing that she has reason to, she is internally inconsistent—she is irrational by her "own lights." But suppose that, unbeknownst to her, arising would do nothing but attract the fatal attentions of the cobra under her bed. Then, in Smith's "full" sense of rationality, her lack of desire is quite rational, because rational in light of the cobra.
The idea that truth is a rational requirement appears again in Smith's argument (1994 , 85–87) for yet another form of internalism, which he dubs "rationalism" (62):

(R)If it is right for agents to Φ in circumstances *C*, then there is a reason for those agents to Φ in *C*.

(Parfit labels this "moral rationalism" [1997, 103].)
Smith's argument for (R) runs as follows, where (1), (2), and (4) are platitudinous, or follow from platitudes, and are thus conceptual truths.

(1)Rational agents will do what they judge themselves morally required to do. (1994, 86–87)
(2)Rational agents will judge *truly*. (1994, 87)

end p.113

Thus

(3) Rational agents will do what they *are* morally required to do. (1994, 85)

But

(4) Rational agents will do what there is (most) reason to do. (1994, 85)

Hence

(R) There is reason to do as morality requires.

We reject (2). Korsgaard (1986 , 11–12) notes that "conclusions drawn from mistaken premises are not *irrational*." We take her to imply that not all false beliefs are irrational,

with which we concur. Statement (1) we might accept: we view rationality as internal consistency (roughly along logical and decision theoretic lines), and perhaps it is internally inconsistent not to do as you think you ought to. On such a construal of rationality, however, (3) and (4) are false (see section 4, where we also deny (R)).
A stronger form of (C2) has it that "we cannot believe that we have a reason to do something without being motivated to do this thing" (Parfit 1997 , 104; he calls this "belief internalism"). The much-discussed claim that there is an internal connection between *moral* beliefs and motivation (e.g., McNaughton 1988 , 23) Parfit (1997 , 105) dubs "moral belief internalism": "We cannot believe some act to be our [moral] duty without being motivated to do it."
Smith's (1994 , 61–62) "practicality requirement on moral judgement,"

(P)If an agent judges that it is right for her to Φ in circumstances *C*, then either she is motivated to Φ in *C* or she is practically irrational,

is a weakening of moral belief internalism. [3] Smith (1994 , 62) argues that (P) is entailed by (R). If a rational agent judges that Φ-ing is right, then, by (R), she judges that she has reason to Φ. But then, by (C2), she desires to Φ.
(R), however, makes no reference to judgments. To yield a valid argument from (C2) to (P), we need not (R), but the following:

(B)If an agent judges that it is right for her to Φ in circumstances *C*, then either she believes that she has a normative reason to Φ in *C* or she is practically irrational.

To yield a valid argument from (R) to (P), we need as additional premises:

(Q1)If an agent judges that it is right for her to Φ in circumstances *C*, then either it is right for her to Φ in *C* or she is practically irrational.

And
end p.114

(Q2)If there is reason for an agent to Φ in *C*, then either she is motivated to Φ in *C* or she is practically irrational.

(Note, however, that (1), which is used as a premise in Smith's argument for (R), yields (P) directly.)

We discuss (Q1) in section 4. (Q2) is false if irrationality is construed as internal inconsistency: suppose you are unaware that the building is on fire; you might be quite consistent in having no desire to leave it despite having good reason to. Those who maintain that we have moral duties while arguing that there is no reason to fulfill them would presumably deny (B) (in section 4 we deny (B) for other reasons). Discussion of the relations between reasons, morality, and motivation raises the issue of amoralism, our next topic.

According to the moral noncognitivist (or moral expressivist), there are no moral facts (see, e.g., McNaughton 1988). It is not that remarks such as "you morally ought to Φ" are false; rather, they lack truth value. And a thoroughgoing expressivist would say the same of "you have reason to Φ." Smith (1994 , 63–66) sees Mackie (1977), on the other hand, as endorsing (R) and arguing from the universal falsity of its consequent to the universal falsity of its antecedent. Brink distinguishes these two forms of skepticism from "amoralist skepticism," which "accepts the existence of moral facts and concedes that we have moral knowledge, and asks why we should care about these facts.Amoralists are the traditional way of representing thiskind of skepticism; the amoralist is someone who recognizes the existence of moral considerations and remains unmoved" (1989, 46). Amoralists here, such as "certain sociopaths" who (apparently) have moral beliefs but lack moral motivations, would be counterinstances to moral belief internalism. (See also McNaughton 1988 , chap. 9.) And Brink (1989 , 46–50) does deny moral belief internalism. [4] But in the quoted passage he sees such amoralists as "representing" not the denial of moral belief internalism, but, rather, (roughly) a challenge to (R): even if we concede that there are moral facts, why should we care about them—that is, is there reason to act on them? He later notes (1989, 48), however, that (R) and moral belief internalism are logically independent of one another.

A Williams-style internalist about reasons, for example, might deny (R) but endorse moral belief internalism: she might claim that even though torture is wrong, the sadist for whom there is no procedurally rational route to a motivation to refrain has no reason to refrain—and he is not a counterinstance to moral belief internalism because he does not view torture as wrong. Note that whether one describes the sadist as amoral is irrelevant here: the amoralist skeptic who is challenging (R) here is the theorist, and she may well be, with good (internal) reasons, the most morally upright of characters.

There are other ways of denying (R) (see Parfit 1997 , 103; sec. 4 below). Parfit sees Foot (1972) as suggesting a view that acknowledges motivation-independent

moral duties but on which "though self-interest provides external reasons, morality does not" (Parfit 1997 , 104).

Smith (1994 , 66–91) responds to Brink (1986 ; 1989 , 46ff.) and Foot (1972) by arguing that they are committed to an inaccurate account of the morally good person. On their accounts, Smith claims (1994 , 72–75, 83–84), the morally good agent manifests a reliable connection between what she believes to be morally required and her motivation to do it because she has a "self-conciously *moral* motive" (1994, 74): a standing motivation to do what is right under the description "do the right thing." And this manifests "a fetish or moral vice" (1994, 75). If we accept (P), however, Smith claims

that we can then explain this reliable connection in all rational agents without appeal to any self-consciously moral motive.

We are not convinced, however, that this claim coheres with Smith's account of desire. He holds that "desiring to Φ [is] having a certain set of dispositions, the disposition to C in conditions C, the disposition to X in conditions C', and so on, where, in order for conditions C and C' to obtain, the subject must have, inter alia, certain other desires, and also certain means-ends beliefs, beliefs concerning Φ-ing by C-ing, Φ-ing by X-ing and so on," and that "the propositional content of a desire [is] determined by its functional role" (1994 , 113–14). But according to (P), whenever a rational agent judges that Φ-ing is right, she is motivated to Φ. Rational agents are disposed to pursue what they take to be the means to their ends. Hence a rational agent is disposed to pursue what she believes to be the means to doing right. And, on Smith's account of desires and their contents, this disposition would appear to be a standing desire to pursue the right, under that description—the very "fetishistic" desire that he denounces.

According to the Humean theory of motivation, beliefs and desires are "distinct existences" (Smith 1994 , 119); thus it is possible for someone to believe that she has (moral) reason to Φ without desiring to Φ. Hence belief and moral belief internalism are false. Smith sees himself as a Humean about motivation (1994 , chap. 4, and 179), but his position is consistent (1994 , 179–180) because he substitutes (C2) and (P), respectively, for belief and moral belief internalism. Thus it is possible, but irrational, for someone to believe that she has (moral) reason to Φ without desiring to Φ. It is only the rational agent who is disposed to do what she believes she has most (moral) reason to do. But is such a disposition distinct from a standing desire to do what one believes one has most (moral) reason to do?

Parfit (1997 , 105) [5] sees such a distinction between disposition and desire as crucial to rejecting the Humean theory of motivation. Parfit's account of this theory is not Smith's, however. According to Parfit, it is the view that "no belief could motivate us unless it is combined with some independent desire," where a desire is independent of a belief if it is not produced by the belief. To reject this, it suffices to maintain that a belief that one has reason to Φ can produce, by itself, a desire to Φ, where the disposition to acquire the desire upon acquiring the belief is not itself a desire. [6] But this latter view counts as Humean by Smith's criteria: "All actions are indeed produced by desires, just as the Humean says; no actions are produced by beliefs alone.Butsome of these desires are themselves produced by the agent's beliefs about the reasons she has" (Smith 1994 , 179; see also 119, 213 n. 3).

Williams's internalism can be interpreted as being consistent with either account of the Humean theory of motivation. Given Williams's concern to tie reasons to motivation, he may well see a belief that one has reason to Φ as accompanied by a desire to Φ, at least in internally consistent agents. But he need not claim that such beliefs produce desires unaided by, say, a desire to do what one believes one has reason to do. Or perhaps the desires can produce the beliefs: your desire to Φ might produce a belief that you have a reason to Φ. *A* has a sudden desire for ice cream, and, as a result, forms the belief that it would be enjoyable. Believing that her enjoyment is a reason, she believes that she has a reason to eat ice cream. (Normatively, Williams can see the desire as fallible evidence of a reason: *A*'s belief that the ice cream would be enjoyable is subject to correction.)

Finally, we turn to Korsgaard's (1986 , 11) "*internalism requirement*": "Practical-reason claims, if they are really to present us with reasons for action, must be capable of motivating rational persons." Smith (1994 , 62, 150) interprets this as tantamount to (W). Parfit (1997 , 107), however, interprets it as consistent with the *denial* of (W). We take it that Parfit is interpreting the phrase "practical-reason claim" as denoting a belief and interprets Korsgaard as meaning something like the following variant on Smith's (C2):

(C2')If a rational agent believes that she has reason to Φ, then she desires to Φ, or could acquire such a desire.

Neither interpretation is unreasonable. Here is our tentative exegesis.
Sometimes Williams states his internalism in a form weaker than (W), as:

(W')*A* has reason to Φ only if she *could* acquire a desire to Φ by rational deliberation from her relevantly corrected *S*.

(See Williams 1981 , 105; 1995a , 35; Parfit 1997 , 116–17.) In section 1, we noted that it is crucial to Williams's internalism that if *A* has reason to Φ, that is *because* she would arrive at a desire to Φ if she deliberated rationally. However, it is consistent with the letter of (W) to claim that reason, not rational motivation, is metaphysically basic: if *A* has a reason to Φ, then *because* of this she would arrive at a desire to Φ if she deliberated rationally (see Hooker 1987 ; Hurley 2001). Similarly, the following are both consistent with (W'):

(W'1)If *A* has a reason to Φ, that is because she could arrive at a desire to Φ if she believed the relevant facts and deliberated rationally.
end p.117

(W'2)If *A* has a reason to Φ, then, because of this, she could arrive at a desire to Φ if she believed the relevant facts and deliberated rationally.

Korsgaard (1986 , n. 17) cites Nagel (1970) as drawing this distinction between "two kinds of internalism," and as arguing for (W'2), so that "investigations into practical reason will yield discoveries about our motivational capacities. Granting that reasons must be capable of motivating usif we then are able to show the existence of reasons, we will have shown something capable of motivating us" (Korsgaard 1986 , 23). Korsgaard (again citing Nagel) sees Kant as accepting (W'2). If the "because" in (W'2) is to have appropriate force, then if reasons exist, they must be capable of motivating us, as opposed to being merely "epiphenomenal" (Korsgaard 1986 , 24). Kant proposes an account of

“pure practical reason” (based on the categorical imperative) and how it can motivate us, and argues that practical reasons do motivate us, if we are rational.
Given that Korsgaard (1986 , 11–12) does not regard false beliefs as irrational, we take it that Korsgaard's Kantian picture portrays the rational agent whose beliefs have been relevantly corrected in Williams's fashion as motivated by the reasons she has, where these reasons accord with the categorical imperative. Furthermore, when rational agents (relevantly corrected) are motivated to Φ, this is because they have reason to Φ: reason is metaphysically basic. Assuming that one way in which reasons motivate us is via belief that we have reasons, this picture is consistent with (C2') (of course, the rational agent can desire to do something she has no reason to do since her beliefs need not be relevantly correct). And assuming that rational deliberation accords with the categorical imperative, this picture is consistent with (W'2).

3. Williams's Internalism: Pro and Con

On the views we have been attributing to them, Williams and Korsgaard both accept (W'). But, as Korsgaard emphasizes, the content of (W') depends on the content of the relevant notion of rationality. Williams is skeptical that rationality includes prudence and morality (Parfit labels a notion of rationality that includes at least prudence as “substantive rationality” (1997, 101)). Korsgaard (1986) agrees with Williams (1995a , 37) that arguments must be given for including morality and prudence within the purview of rational motivation and agency. But, first, why do we not also need arguments to show that means/ends reasoning belongs
end p.118

within this purview (24)? And second, Korsgaard sees Kant as providing arguments for the inclusion of morality within it. (See also Velleman 2000 , 173.) This argument accuses the immoral of irrationality. But one can be internally consistent and yet immoral (and imprudent): for instance, violations of the categorical imperative need not violate logic or decision theory. Thus on the notion of rationality as internal consistency, this latter argument fails. (Of course, Kantians do not see rationality as mere internal consistency: they have something much more substantial in mind.)
We are more sympathetic to Korsgaard's challenge to (W'1). We agree with her that reason is metaphysically basic. However, we deny (W') (and hence deny both (W'1) and (W'2)), again because we see rationality as a matter of internal consistency: since an agent need not manifest inconsistency in lacking any “interest in preserving his health” (Williams 1981 , 106), he need not manifest inconsistency in being incapable of acquiring a desire to take the medicine, despite having reason to and having relevantly correct beliefs in Williams's sense.
Williams's restrictions on belief correction raise a complaint, related to Korsgaard's general thrust, that can be leveled by certain externalists who see morality and prudence as providing reasons independently of considerations of rationality. These externalists claim it a fact that we have reason to care about our health and see Williams as begging

the question by presuming that such normative facts are not to be listed as potential sources of error in an agent's *S*. Williams allows us to add to the agent's *S* the belief that she needs the medicine, but if she does not want to recover and cannot be persuaded otherwise, [7] then, by Williams's lights, she has no reason to take it (1981 , 105–6). We are not allowed to add to the agent's *S* the belief that she has reason to recover.

Williams responds to challenges concerning the content of rationality thus:

> any rational deliberative agent has in his *S* a general interest in being factually and rationally correctly informed.[Thus] on the internalist view there isa reason for writingthe requirements of correct information and reasoning into the notion of a sound deliberative route, but [there is] not a similar reason to write in the requirements of prudence and morality. Somebody may say that every rational deliberator is committed to constraints of morality as much as to the requirements of truth or sound reasoning. But if this is so, then the constraints of morality are part of everybody's *S*, and every correct moral reason will be an internal reason. But there has to be an argument for that conclusion. Someone who claims that the constraints of morality are themselves built into the notion of what it is to be a rational deliberator cannot get that conclusion for nothing. (1995a, 37)

As Williams notes (1995a , 44 n. 3), he and Korsgaard (1986) agree that argument is required to establish that rationality incorporates morality. But they disagree (we take it) over whether such arguments are to be found in, say, Kant, and over the issue of whether argument is required to establish that "any rational delib

erative agent has in his *S* a general interest in being factually and rationally correctly informed."

Williams's response here does not address, however, our externalist who wonders why the fact that we have moral and prudential reasons is excluded from the realm of correct information. But Williams has more to say, to which we shall turn after addressing Smith's and McDowell's complaints against him.

Smith acknowledges (1994 , 156) the similarity of his account of reasons to that of Williams, but he has two complaints. First (1994 , 158–61), Williams ignores the role of rational deliberation in "systematically justifying" desires. By analogy with Rawlsian reflective equilibrium, Smith sees rational deliberation as a route to unifying and modifying our desires. We are skeptical. Deliberation might result in the realization that one would be better off without a desire for a certain end. It might also eliminate motivation to pursue means to that end. Yet the desire might persist. Furthermore, does rationality require systematically unified desires? When we speak of internal consistency among propositional attitudes, we do not see this as applying to desires. It is not inconsistent, in our sense, to desire to remain in bed and to desire to get up (although it is inconsistent to intend to do both).

Smith's (1994 , 164–74) second complaint concerns Williams's "relativism." On Williams's account of reasons it is conceptually possible that, although *A* in circumstance *C* has a reason to Φ, *B* may not. This because their *S*s may differ, even after relevant correction. Smith, however, maintains that it is part of our conception of a reason that if *A* and *B* face the same circumstances, they have the same (normative) reasons. Thus, given (W), "convergence in the hypothetical desires of fully rational creatures is required for

the truth of normative reason claims" (Smith 1994 , 173). Smith claims that it is a "substantive" question whether "such a convergence is forthcoming," and hence a substantive question whether there are any reasons. He sees his argument with Williams as over the conceptual issue of whether reasons require convergence, not over the substantive issue of whether there are reasons.

It is unclear to what extent Smith tackles Williams head-on here, however. For instance, Smith (1994 , 171–72) accuses Williams of adhering to a counterintuitively relative account of reasons according to which we cannot disagree over whether you have reason to take a holiday. Relative to your *S*, you do, but relative to mine, you do not: were I facing your circumstances, unlike you I would not arrive at a desire to take a holiday. But Williams would say that what I would do in your circumstances is irrelevant. There is a nonrelative fact of the matter over which we can disagree, namely whether you would arrive, upon rational deliberation from your current *S* relevantly corrected, at a desire to take a holiday. If you would not, then you have no reason to take a holiday.

For Williams, whether an agent has moral and prudential reasons depends on whether she would arrive at the relevant desires were she to deliberate ration

ally from her current *S* relevantly corrected. Since some of us do so arrive at such desires, some of us do have such reasons. For Smith, by contrast, like Korsgaard, whether an agent has prudential and moral reasons depends on whether prudential and moral desires are required by rationality. If such desires are not so required, then none of us have such reasons. But whereas Korsgaard appeals to Kantian arguments concerning the requirements of rationality, Smith (1994 , 173–74) sees it as a matter of more general "convers[ations] and argu[ments] about the reasons we have." We have moral and prudential reasons if and only if these conversations and arguments converge upon that conclusion. Our response, as it was to Korsgaard, is to note that one can lack moral and prudential desires without being internally inconsistent. We agree that there are moral and prudential reasons, and that these are nonrelative in the sense that they can be independent of the contents of an agent's *S*. But we do not tie reasons to rationality as closely as Smith or Korsgaard.

One thread in McDowell's (1995) argument against Williams runs as follows. According to the externalist, on Williams's account, the immoral and imprudent are irrational (1981, 110), regardless of whether they have internal reasons to be moral and prudent. And thus, on the externalist account, the immoral and imprudent may need further *rational* input beyond mere rational procedure from a relevantly corrected *S*. But McDowell points out that the further input need not be rational. All the externalist claims is that the immoral and imprudent are deliberating incorrectly, and a "transition *to* deliberating correctly [need not be] effected *by* deliberating correctly; effecting the transition may need some non-rational alteration like conversion" (McDowell 1995 , 78). The immoral and imprudent suffer a deliberative flaw, but it need not be one of irrationality.

Williams acknowledges (1995b , 192) that the externalist need not accuse the immoral and imprudent of irrationality. But he challenges McDowell to give an externalist account of "deliberating correctly" that will furnish an account of reasons that "represents a statement about *A*'s reasons as a distinctive kind of statement about, distinctively, *A*"

(194). McDowell suggests [8] that "deliberating correctly [might be a matter of] giving all relevant considerations the force they are credited with in a correct picture of one's practical predicament. This yields a sense in which to believe an external reason statement is, as Williams indeed suggests it must be (1981, 109), to believe that if the agent deliberated correctly he would be motivatedin the direction in which the reason points" (McDowell 1995 , 78). Williams asks, Who deliberates correctly? McDowell (1995 , 74) speaks of the person who has "been properly brought up," that is, an Aristotelian *phronimos* (Williams 1995b , 189ff.). Thus, apparently, you have most reason to do what the *phronimos* would do in your circumstance. Williams (1995b , 190) responds that this fails to take into account an important feature of your circumstance, namely the fact that you are *not* a *phronimos*.

In part, Williams seems to apply a "reason implies can" principle: if you have reason to Φ, you must be capable of Φ-ing, and you may not be capable of doing as the *phronimos* does. The externalist needs an account of reasons that does not require, for instance, "moral weight-lifting" (Williams 1995b , 190) beyond your capability. Furthermore, Williams asks (1981 , 110; 1995a , 39–40), why, when there is no rational route to motivation, does the externalist phrase her criticism of an agent in terms of *reasons*, when there are plenty of other criticisms to be made (such as complaints concerning the agent's nastiness, cruelty, selfishness, imprudence, brutality, and so on)?

In our externalist view, however, to make one of these latter complaints is to give a reason against the act in question. To say that an act you are contemplating is cruel is to give a reason against doing it—distinct from a reason of, say, imprudence (see also Scanlon 1998 , 367). On the internalist picture, its cruelty is a reason against your action only if you can be rationally brought to be motivated not to do it. According to externalism this need not be—for the externalist, reasons carry normative force, but this should not be confused with motivational force (as Parfit 1997 , 111–12, emphasizes). Williams claims (1995a , 39) that "internalist theory explains how it is that the agent's accepting the truth of `There is reason for you to Φ' could lead to his so acting, and the reason would thus explain the action. It is obvious on the internalist view how this works." But it is the *belief* internalist who claims that you cannot believe you have reason to Φ without being motivated to Φ (Parfit 1997 , 104). And it is open to the reasons externalist to be a belief internalist (as it is to the reasons internalist to deny belief internalism).

Suppose you point out to someone that what he is about to do is cruel. This might suffice to prevent him doing it. But you might have to explain why its cruelty is a reason against doing it. If this succeeds in preventing the action, neither reasons internalism nor reasons externalism explains this. Either theory needs augmentation to accomplish this, and neither is more easily augmented than the other.

Williams has another tack with the thought: "If it is true that *A* has a reason to Φ, then it must be possible that he should Φ for that reason" (1995a , 39). On the externalist account, a sadist has a reason to refrain from cruel action even if there is no procedurally rational route to his being so motivated. But is the sadist necessarily so, or is there some possible world in which he refrains because it is cruel? Furthermore it may in some psychological sense be impossible to correct someone's *S* or get them to be procedurally rational (cf. n. 7).

What of the *phronimos?* One approach is to take her as metaphysically basic and derive an account of reasons: *A* has reason to Φ if and only if, and because, the *phronimos* would be motivated to Φ in *A*'s circumstance. The converse approach has it that the *phronimos* would be motivated to Φ in *A*'s circumstance if and only if, and because, *A* has reason to Φ. We reject both approaches. We see reasons as metaphysically basic. But we agree with Williams that the constitution
end p.122

of the individual has a role in determining what she has reason to do. We doubt it makes sense to ask, say, whether the *phronimos* enjoys fell-walking. But you might, in which case you have reason to do it (unless, perhaps, you are the Puritan below).
What of the relation between motivation, rationality, and reasons on our externalist view? Suppose it a fact that

(E) *A* would enjoy fell-walking.

Then there is a further fact: the fact that

(F) (E) gives *A* a reason to fell-walk.

(It is (F) that is a normative fact here [cf. Parfit 1997 , 124].)
A might be unaware of either fact or both. *A* is not irrational in failing to appreciate (E): she is merely ignorant. But suppose *A* is a Puritan who denies that her enjoyment is ever a reason. Is *A* irrational? We doubt it. Indeed, it may be that reasons are sufficiently personal that *A* is not only rational, but correct. The crucial point is that we, unlike Williams, see the issue of whether (F) is true as independent of the availability, or otherwise, of a procedurally rational route to get *A* from her relevantly corrected *S*, which (we take it) includes (E), to a desire to fell-walk.
Williams is puzzled over the meaning of external reasons claims (1981 , 110–11; 1995a , 39–40). But why is he any less puzzled over the meaning of internal reasons claims? Contrast a Puritan with an ordinary person, both of whom know they would enjoy fell-walking, and both of whom desire to do so. The Puritan is busy ridding herself of the desire. The ordinary agent is contemplating whether to head for the fells. The difference between them is that the ordinary person believes that her enjoyment is a reason here, whereas the Puritan does not. But what is the content of this difference on Williams's view? Williams agrees (1995b , 187–88) that an agent can arrive by correct deliberation at the belief that she has a reason to Φ. But he (sensibly) denies that this is to arrive by correct deliberation at the belief that if she deliberated correctly she would be motivated to Φ*;* hence, he argues, "*A* has reason to Φ" is not equivalent to "If *A* deliberated correctly, he would be motivated to Φ."

The externalist can see reasons as personal: the Puritan's and the ordinary person's beliefs about their respective reasons might both be true. On the other hand, perhaps the Puritan does have a reason to go fell-walking. Williams gives no positive view of the content of beliefs about reasons; thus his account of a debate with the Puritan is no clearer than an externalist account that allows us, in trying to convince the Puritan that she has reason to go fell-walking, to argue that fell-walking is a harmless pursuit; exercise, relaxation, and the appreciation of beauty are beneficial; and she will return rejuvenated to her Puritan endeavors.
end p.123

Finally, the Puritan's belief that she has no reason to enjoy herself, in light of which her desire to fell-walk is irrational, is an element of her *S*. Williams does not allow, in determining the Puritan's internal reasons, correction of this belief (otherwise we could correct beliefs about such matters as whether one has reason to care about one's own further future). But *why* is it the desire that is subject to normative correction and not the belief? We have returned to the question-begging complaint.
This completes our discussion of Williams's internalism. We turn finally to the issue of what one ought to do.

4. Reasons, Oughts, and Rationality

One simple proposal [9] is:

(O) "*A* ought to Φ" is equivalent to "*A* has most reason to Φ (and Φ-ing is an option)."

We label it (O) because it is, in Prichard's (1932) sense, an objective view of ought. Here is one of his examples against such views: "`Ought we to stop, or at least slow down, in a car, before entering a main road?' If the objective view be right, (1) there will be a duty to slow down only if in fact there is traffic; (2) we shall be entitled only to think it likely—in varying degrees on different occasions—that we are bound to slow down; and (3) if afterwards we find no traffic, we ought to conclude that our opinion that we were bound to slow down was mistaken" (1949, 29). Given that there is no traffic, what you have most reason [10] to do is proceed apace—regardless of your state of knowledge (this is so even on an internal view of reasons, since your *S* would be corrected to incorporate the true belief that there is no traffic). But surely, if you are uncertain about the traffic, what you ought to do, counter to (O)'s rash prescription, is slow down.
In defense of an objective view, one might suppose that there is an objective probability that a car is coming, and that the agent should slow down because this probability is greater than zero. Prichard rejects this sort of account because:
there are no such things as probabilities in nature. There cannot, e.g., be such a thing as the probability that someone has fainted, since either he has fainted or he has not. No

doubt it is extremely difficult to formulate the precise nature of the fact which we express, for instance, by the statement: “X has probably fainted.” But at least we must allow that, whatever its precise nature may be, the fact must consist in our mind's being in a certain state or condition. And,
end p.124

once this is realized, it becomes obvious that most of our ordinary thought involves the subjective view. (1949, 30)

This passage is suggestive of a Bayesian view according to which all probabilities are subjective. [11] One key feature of Bayesianism is the idea that subjective probability is a property of the believer rather than of the object of belief: if your degree of belief that a tossed coin will come up heads is one half, this is a property of you rather than the coin. On pure Bayesianism, probability assignments held by an agent are criticizable if and only if they collectively violate Kolmogorov's (1933) axioms. [12] This comports with the idea of rationality as internal consistency: internal consistency of subjective probabilities is conformity to Kolmogorov's axioms. (Such internal consistency is, more or less, what Joyce [chap. 8, this volume] refers to as “probabilistic consistency.”)

Decision theory (see Joyce, chap. 8, and Dreier, chap. 9, this volume), according to which the agent ought to act so as to maximize her subjective expected utility, where the latter is a function of her utilities for outcomes and subjective probabilities, can be viewed as laying down criteria for internal practical consistency more generally (see Jeffrey 1983 ; Rawling 2003). Although he invokes no technical apparatus (he was presumably unaware of Ramsey's (1931) groundbreaking work), Prichard's subjective ought is not far from incorporating decision-theoretic considerations (see, e.g., 1949, 26). He is, however, concerned with moral duty, and he seems to suppose that the agent knows which circumstances are moral reasons for which actions. The difficulty on which he focuses is that the agent might not know the circumstances. Duty, he claims, is then determined by the agent's subjective probabilities concerning them. Thus, in the case of the junction, the agent knows that traffic is a reason to stop, but that in its absence she has no reason to (moral or otherwise, in this case). Even if, however (unbeknownst to her), there is no traffic, she ought to slow because her subjective probability that traffic is present is sufficiently high.

On our view, like Prichard's, (O) is false. However, we suspect that there is a reading of “ought” that is less subjective than Prichard's. We disagree, for instance, with Prichard's view that if a “would-be torturer [were] in a very high degree confident that torturing, and torturing only, would save the heretic, he would be bound to inflict the torture” (1949, 30). Such a “high degree” of confidence strikes us as unreasonable, and hence the would-be torturer ought not to torture. Admittedly, the notion of a reasonable subjective probability is vague, and we shall do no more here, we fear, than appeal to intuitions in taking the notion as given. The notion of a reasonable subjective probability, requiring more than mere conformity to Kolmogorov's axioms, is to be found in, for example, Ramsey (1931 , sec. 5). Ramsey links such a notion to the idea of a “useful [mental] habit”; we shall not commit ourselves that far here, however.

Our claim is that what a person ought to do depends upon what it is rea
end p.125

sonable for her to surmise about her circumstances, and what it is reasonable for her to suppose concerning which circumstances are reasons for which actions. One ought to slow at a traffic junction not because one happens to have a sufficiently high subjective probability that traffic is present, but because such a subjective probability is reasonable. Someone who fails to slow because he has, despite lacking the dispositive evidence required to support such a verdict, a zero such subjective probability, does not do as he ought. Admittedly, if there is no traffic, he has no reason to slow—but this just shows that he may have no reason to do the moral and/or prudent thing (we reject [R] from sec. 2 above, and its prudential variant).

One of Prichard's worries about the subjective view is "that it represents the duty of doing some action as depending not on the fact that the action would have a certain character if we were to do it, but on our thinking it likely that it would. And to maintain this seems impossible" (1949, 31). Ultimately he resolves this difficulty by arguing that what we are obligated to do is will [13] something, and that subjective probabilities "enter into the character of [the willing]to will *X*, thinking it likely to produce *Y*, is one willing, and to will *X*, thinking it unlikely to produce *Y*, or to will *X*, not thinking of *Y* at all, is another" (1949, [2], 38). However, we do not see how this resolution alleviates Prichard's related worry (1949, 26) that, although it seems that the ground of an omniscient's obligation to Φ is the fact that Φ-ing would have some outcome, yet on the subjective view the ground of her obligation to Φ is not the fact that Φ-ing would have a certain outcome, but the fact that she knows this.

In our view, however, it is the fact that one has a *reason* to Φ that is grounded in the outcome of Φ-ing. Whether I *ought* to (attempt to) Φ depends upon whether a reasonable person in my circumstance would (attempt to) Φ. If I am omniscient, this is part of my circumstance, and a reasonable person in this circumstance would have subjective probabilities of ones and zeroes in truths and falsehoods respectively. But what an omniscient ought to do remains grounded in what a reasonable person would do in her circumstance. When we engage in practical reflection, we try to determine what we have most reason to do. But ignorance obtrudes, and what we ought to do takes reasonable account of this, forcing a conceptual separation from what we have most reason to do, and often forcing a separation in output. In the case of omniscients, the two always coincide in output but maintain separate grounds.

The circumstance of an ordinary agent also includes the fact that she is not a *phronimos*, and just as what she has reason to do must take this into account, so must what she can reasonably be expected—and hence ought—to do. A person weak of will with respect to certain damaging temptations, for example, has reason to avoid them, and she ought, ceteris paribus, to take reasonable precautions to do so. Both oughts and (external) reasons, then, take into account various features of the agent. But oughts are more sensitive to her epistemic and motivational capabilities: there may be a weak "reason implies can" principle (does an agent have reason to do something that is impossible for her?), but the "ought implies can" principle is stronger.

Prichard is largely concerned with an ought of morality, whereas we are pursuing the notion of what an agent ought to do all things considered. Reasons and oughts can be divided into various categories (moral, prudential, etc.). And, on one reading of "ought,"

moral and other oughts can compete in the way that moral and other reasons can. It might be that, say, you morally ought to Φ, [14] in the sense that a reasonable agent in your circumstances would see Φ-ing as favored when looked at only from the moral point of view, but if the total considerations (from the reasonable perspective) favor (in roughly decision-theoretic fashion) not Φ-ing, then, all things considered, you ought not to Φ.

It should be noted that, despite the subjective [15] components in Prichard's view of ought, and ours, both views are in other ways objective. First, on both accounts there is a straightforward fact of the matter as to what an agent ought to do. Second, on both accounts an agent can be mistaken about what she ought to do. Prichard correctly sees "as viciousthe statement `*Doing so-and-so* would *be* right because *I think* it would be right'" (1949, 37). For him, what one ought to do is subjective only to the extent of depending on one's view of the circumstances—one's normative beliefs concerning what one ought to do given that view seem to play no role. It is the latter objectivity that leaves room for error. On our view, what one ought to do is dependent both upon how a reasonable person would perceive one's circumstances, and on what she would believe she ought to do given that perception. For us, error is the result of unreasonableness.

Brink sees moral realism as both committed to moral facts and opposed to relativism (1989 , 14). He argues for "a single true morality" (1989 , 53, 77). This we regard as too strong a criterion, at least on one reading. Brink reconstructs and challenges an argument from (R) and (roughly) Williams-style internalism to the relativist conclusion that "there will be different moral requirements that apply to different people in virtue of their different motivational sets" (1989 , 52). We reject the premises, but we have some sympathy with the conclusion: what an agent ought to do is sensitive to some degree to her motivational capacities. We (and Prichard) deny that "one can have a moral obligation to [Φ] only if [Φ-ing] would contribute to the satisfaction of one's desires" (Brink 1989 , 52). But our view is relativist in incorporating an "ought implies can" principle. This is consistent, however, with the appropriate degree of realism: it is a fact, for instance, that Hitler (cf. Brink 1989 , 55–56) ought not to have done as he did (although to say that he was merely unreasonable is, we admit, an understatement).

A reasonable person may be unable to determine what she has reason to do, but she can determine what she ought to do. [16] Someone can fail to realize what she ought to do, however, and yet be perfectly rational (we deny (3) from section

end p.127

2): internally consistent agents can not only fail to see what they have reason to do, they can also be unreasonable. (See also Scanlon 1998 , chap. 1, and the conclusion to Hooker and Streumer, chap. 4, this volume.)

We conclude by addressing (4), (B), and (Q1) from section 2.

We deny (4). An agent might fail to do what there is most reason to do and yet be rational (and reasonable): it need not be irrational (nor unreasonable) to fail to appreciate your circumstance, nor to fail to see that a particular circumstance is a reason to perform a certain act.

We deny (B). An agent might judge that it is right to Φ, and yet be internally consistent in lacking the belief that she has reason to: she judges, for instance, that she ought to slow

down, while rationally acknowledging that there is likely no traffic, and hence likely no reason to slow.
Finally, we deny (Q1) as it stands, but we accept it with "unreasonable" substituted for "irrational," thus:

If an agent judges that it is right for her to Φ in circumstances *C*, then either it is right for her to Φ in *C* or she is practically unreasonable.

APPENDIX OF PROPOSITIONS

(C2) If an agent believes that she has a normative reason to Φ, then she rationally should desire to Φ.
(W) *A* has normative reason to Φ in *C* only if *A* would desire to Φ in *C* if fully rational.
(R) If it is right for agents to Φ in circumstances *C*, then there is a reason for those agents to Φ in *C*.
(Alternatively: There is reason to do as morality requires.)
(1) Rational agents will do what they judge themselves morally required to do.
(2) Rational agents will judge *truly*.
(3) Rational agents will do what they *are* morally required to do.
(4) Rational agents will do what there is (most) reason to do.
(P) If an agent judges that it is right for her to Φ in circumstances *C*, then either she is motivated to Φ in *C* or she is practically irrational.
(B) If an agent judges that it is right for her to Φ in circumstances *C*, then either she believes that she has a normative reason to Φ in *C* or she is practically irrational.
(Q1) If an agent judges that it is right for her to Φ in circumstances *C*, then either it is right for her to Φ in *C* or she is practically irrational.
end p.128

(Q2) If there is reason for an agent to Φ in *C*, then either she is motivated to Φ in *C* or she is practically irrational.
(C2') If a rational agent believes that she has reason to Φ, then she desires to Φ, or could acquire such a desire.
(W') *A* has reason to Φ only if she *could* acquire a desire to Φ by rational deliberation from her relevantly corrected *S*.
(W'1) If *A* has a reason to Φ, that is because she could arrive at a desire to Φ if she believed the relevant facts and deliberated rationally.
(W'2) If *A* has a reason to Φ, then, because of this, she could arrive at a desire to Φ if she believed the relevant facts and deliberated rationally.
(E) *A* would enjoy fell-walking.
(F) (E) gives *A* a reason to fell-walk.
(O) "*A* ought to Φ" is equivalent to "*A* has most reason to Φ (and Φ-ing is an option)."

NOTES

We thank Eve Garrard, Josh Gert, Brad Hooker, Al Mele, and the members of a 2001 graduate seminar on reasons at Florida State University for their helpful comments.

1. Smith (1994 , 62) attributes (roughly) (W) to Korsgaard (1986 —see her "internalism requirement," 11; see also below), who is in turn responding to Williams. While acknowledging the similarity between his account and that of Williams (Smith 1994 , 156), Smith distinguishes the two (158–61, 164–74; see also sec. 3 below).

2. Thanks to Robert Dunn for pointing this out to us.

3. On the assumption that judging Φ to be right is tantamount to believing Φ to be your moral duty—an assumption that noncognitivists (expressivists) deny (see McNaughton 1988 ; Smith 1994 , chap. 2).

4. Smith also, it would seem, could deny moral belief internalism—he endorses only (P). Thus he could declare that it is possible but irrational to believe Φ-ing to be morally required and yet fail to be motivated to Φ. He (1994, 66–71) does not take this line, however, but appears to side with the moral belief internalist in denying the possibility of an amoralist who believes she is morally required to Φ and yet lacks any motivation to do so.

5. Citing Nagel 1970 ; Dancy 1993 and 1995 ; and Snare 1991 .

6. Parfit says, "Such a belief could not all by itself cause us to have this desire, since we would have to be *such that*, if we came to have this belief, that would cause us to have this desire. But this disposition may not itself be a desire." In this sense of "cause," my lighting the match could not all by itself cause an explosion, because the world would have to be such that my lighting the match would cause an explosion. This seems to imply that the world being a certain way (the background) is part of the cause, since the lighting is not a cause "all by itself." But do we then want to say that

end p.129

lighting the match in the presence of the background conditions could not all by itself cause an explosion, because the world has to be such that lighting the match in the presence of the background conditions would cause an explosion? We do not need, however, to endorse Parfit's way of talking about causation in order to make the relevant points.

7. Williams might also be accused of begging the question on the issue of persuasion. He allows for certain factual corrections in your *S* and looks to desires arrived at in procedurally rational fashion. But what if you want to recover, but cannot be persuaded that you need the medicine or are obdurate in your procedural irrationality?

8. With credit to Hooker (1987), who credits in turn Robert Gay; see McDowell 1995 , 84, n. 15.

9. This might, for example, be read into Parfit 1997 , 111–12, if reason is external.

10. Prichard does not put this matter in terms of reasons, and Dancy (2000 , 56) would presumably reject our interpretative license. From our perspective, Dancy (2000 , chap. 3) does not distinguish sufficiently between what one has most reason to do and what one ought to do.

11. Ramsey (1931) and de Finetti (1930 , 1937) independently "crystallized" (Edwards, Lindman, and Savage 1963) the notion of subjective probability.
12. See de Finetti 1972 , 67–69; Jeffrey 1983 , 80; Resnik 1987 , 47–54; Rawling 2003 .
13. Prichard argues that all I am capable of is, say, willing that I move my hand—my moving my hand is an effect (1949, 32–34; see also Prichard's "Acting, Willing, Desiring" [1949, 187–98], and Davis 1979 , 41). The following is often true, however: if I were to will that I Φ, my Φ-ing would be the effect. This is what it is for Φ-ing to be an option, on our usage. And it is often reasonable, when Φ-ing is an option, to expect an agent to see that it is an option—to see that if she were to will that she Φ, her Φ-ing would be the effect.
14. Currently we are toying with the following classification. Where *P* abbreviates "considerations pro your Φ-ing" and *C* abbreviates "considerations con your Φ-ing" (for the purposes of this note, all considerations are as seen from "the reasonable perspective"):
You morally ought to Φ if and only if the moral *P* outweigh the moral *C*.
You have a moral duty to Φ (not to Φ would be morally wrong) if and only if the moral *P* significantly outweigh the total *C*.
Your Φ-ing would be supererogatory only if: (1) there are moral *P*, but they do not outweigh the nonmoral *C;* (2) the moral *C* are nil or insignificant; and (3) it is not the case that the nonmoral *C* significantly outweigh the moral *P*.
This allows for small acts of kindness and grand acts of self-sacrifice both to be supererogatory. And it can be supererogatory for one to Φ, even though reasonable people can disagree as to whether you ought to have done so, all things considered (there is a degree of vagueness concerning the weighing of considerations, so that although reasonable parties might agree that the moral *P* do not outweigh the nonmoral *C*, they might disagree over whether the latter outweigh the former). But if the nonmoral *C* outweigh the moral *P* by a wide enough margin, the account allows that Φ-ing is mad rather
end p.130

than supererogatory—consider the case of throwing yourself on a grenade in order to muffle it and thereby mitigate the painful assault on bystanders' ears.
Also, there is conceptual space for acts that one ought to do all things considered, and largely on moral grounds, but which are instances of neither duty nor supererogation—returning a greeting might be a case in point.
Note that the conditions for supererogation are not sufficient. We consider here neither the motivational element involved in supererogation nor issues concerning the presence of nonmoral considerations in favor of the supererogatory act. (It might be supererogatory to Φ even though there are significant nonmoral considerations in favour of Φ-ing, provided that Φ-ing is performed from the appropriate motive.)
15. See Dancy (2000 , chap. 3) for further discussion of objectivity and subjectivity in this area.
16. Note that often there are several admissible options from among which the agent ought to pick. Similarly there may be no unique act she has most reason to perform.

Chapter 8
BAYESIANISM
James M. Joyce

Bayesianism provides a unified theory of epistemic and practical rationality based on the *principle of mathematical expectation*. [1] In its epistemic guise it requires believers to obey the *laws of probability*. In its practical guise it asks agents to maximize *subjective expected utility* (see also Dreier, chap. 9, and Bicchieri, chap. 10, this volume).
This essay will be concerned primarily with Bayesian epistemology. Its first section defends the view that beliefs come in varying grades of strength. Section 2 explores the Bayesian requirement of *probabilistic consistency* for graded beliefs. Section 3 explicates Bayesian confirmation theory. Section 4 discusses the thesis that rational belief change proceeds via *conditioning*. Section 5 addresses the charge that Bayesianism engenders an untenable *subjectivism*.

1. Graded Beliefs and Conditional Beliefs

Bayesians maintain that any adequate epistemology must recognize that beliefs come in varying *gradations of strength*. They seek to replace the *categorical* notion of belief as an all-or-nothing attitude of accepting a proposition as true with a graded conception of belief as *level of confidence*. In general, a person's
end p.132

level of confidence in a proposition *X* will correspond to the extent to which she is disposed to presuppose *X*'s truth in her theoretical and practical reasoning. [2]
There are two compelling reasons for thinking that opinions vary in strength. First, this is needed to make sense of decision making. To explain why Smith will bet on Stewball at 4-to-1 odds but not at even odds, we must suppose that she is not sure that Stewball will lose but is more confident that he will lose than that he will win. Second, since evidence comes in a wide variety of types and strengths, a person will be able to proportion her beliefs to her evidence only if her beliefs come in gradations. Someone who wants to know whether a coin will land heads when tossed might have as evidence any proposition of the form "the coin was fairly tossed a thousand times and *n* heads came up." Each value of *n* calls for a different doxastic attitude.
People also have graded *conditional* beliefs that express their confidence in the truth of propositions *on the supposition that other propositions are facts*. If a person has determinate unconditional beliefs in *Y* and *X* & *Y*, and is not certain of *Y*, then her belief in *X* conditional on *Y* will be a function of her unconditional beliefs in *Y* and *X* & *Y*. That said, a person may hold a definite belief about *X* conditional on *Y* even when she has no determinate unconditional beliefs for *X* & *Y* and *Y*, or when she is certain that *Y* does *not* obtain. Conditional beliefs are best seen as sui generis judgments that cannot be *reduced* to unconditional beliefs. [3]

A crucial challenge for Bayesians is to explain how strengths of beliefs can be measured. It is often said that Bayesianism is committed to the existence of sharp, numerical *degrees of belief*. This requires each believer to have a *confidence measure* that, for each proposition X and condition Y, specifies a number $c_Y(X)$ that gauges her level of confidence in X conditional on Y. For fixed Y, $c_Y(\,)$ captures the person's degrees of belief on the supposition Y, and her unconditional beliefs are given by $c(\,) = c_T(\,)$ where T is any truth of logic. By convention, $c(X) = 1$ indicates complete certainty in X's truth, while $c(X) = 0$ indicates complete incredulity.

The idea that people have numerically sharp degrees of belief has been widely criticized both for psychological implausibility and because it requires people to hold opinions far more definite than their evidence warrants. [4] As a result, most Bayesians now grant that few graded beliefs can be precisely quantified. Instead, beliefs are represented variously by interval-valued probabilities, convex sets of confidence measures, or confidence orderings. The most general approach is to characterize a person's opinions using a system of descriptive constraints that might include any of the following sorts of statements:

- She is more confident in X than in Y.
- She believes Z to at least degree 1 3 but at most degree 3 4 .

end p.133

- She believes X conditional on Y more strongly than she believes Z conditional on W.

The person's graded beliefs can then be represented by the *setCon* of all confidence measures that satisfy the constraints. For this system, *Con* contains all measures: $c(X) > c(Y)$; $3/4 > c(Z) > 1/3$; $c_Y(X) > c_W(Z)$. In ideal cases the constraints will be so detailed that *Con* contains a single measure, but it will more commonly contain many measures. Facts about the person's opinions are given by properties that all *Con*'s elements share, for example, she can be said to believe X to degree x only when $c(X) = x$ for *everycCon*.

We can now state the first core tenet of Bayesian epistemology.

Thesis of Graded Belief. Any adequate epistemology must recognize that opinions come in varying *gradations of strength*. A person's graded beliefs can be represented using a set *Con* of confidence measures. Facts about her beliefs correspond to properties shared by all functions in *Con*.

2. The Requirement of Probabilistic Consistency

The second core tenet of Bayesian epistemology requires that rational beliefs be consistent with the laws of probability. A probability function P assigns non-negative real numbers to propositions in such a way that

Normalization. $P(\mathrm{T}) = 1$ for T any truth of logic.

Additivity. [5] $P(XY) = P(X) + P(Y)$ if X and Y are logically incompatible.
A *conditional probability* assigns numbers to pairs of propositions X/Y in such a way that
Probability. $P(/Y)$ is a probability for every Y with $P(Y) > 0$.
Conditional Normalization. $P(Y/Y) = 1$.
Conditioning. $P(X/Y \,\&\, Z) \times P(Y/Z) = P(X \,\&\, Y/Z)$. [6]
end p.134

Hereafter, "the laws of probability" will denote these five requirements.
Here are two useful consequences of these laws:
Logical Consequence. If X entails Y, then $P(Y)$;eg $P(X)$.
Bayes's Theorem. [7] $P(X/Y) = [P(X) \div P(Y)] \times P(Y/X)$.
Logical Consequence ensures that probabilistic reasoning respects deductive logic.
Bayes's Theorem relates the "direct" probability of X conditional on Y to the ratio of the unconditional probabilities of X and Y, and the so-called *inverse probability* (or "likelihood") of Y conditional on X.
Here is the basic Bayesian requirement of epistemic rationality:
Probabilistic Consistency. A rational subject's beliefs must conform to the laws of probability in the sense that *at least one* confidence measure that represents her beliefs must *be* a probability.
In other words, there must be a c in *Con* and a conditional probability such that $c_{Y}(X) = P(X/Y)$ whenever $P(X/Y)$ is defined.
The most common Bayesian rationale given for probabilistic consistency is the famous Dutch Book Argument (DBA) of Frank Ramsey (1931) and Bruno de Finetti ([1937] 1964). This argument purports to show that anyone whose beliefs violate the laws of probability is *practically irrational*. In broadest outline the reasoning runs thus:
Coherence. A practically rational agent will never freely perform any action when another act is certain to leave her better off in *all* possible circumstances.
Belief/Desire Psychology. A practically rational agent will always act in ways that she estimates will best satisfy her desires.
The EU-Thesis. A practically rational agent will estimate that an act best satisfies her desires if and only if that act maximizes her *subjective expected utility*.
Dutch Book Theorem. An agent who tries to maximize her subjective expected utility using beliefs that violate the laws of probability will freely perform an act that is sure to leave her worse off than some alternative act would in all circumstances.
end p.135

Therefore, it is *practically irrational* to hold beliefs that violate the laws of probability.
While the DBA's first premise has generated little controversy, its second is often dismissed as bad psychology. A vast body of experimental evidence shows that, in addition to beliefs and desires, actions are affected by emotions, habits, decision-making heuristics, and judgmental biases. [8] This is all undeniable, but those who see it as a problem for the DBA are confused about the Bayesian enterprise. Bayesians have always made it clear that they are offering a *normative* theory of *rational* behavior, *not* an empirical theory of actual behavior. [9] Emotions, habits, and so on do *cause* actions, but

the DBA does *not* rely on the belief/desire model as a causal theory of action. Rather, the model serves to determine which actions an agent has *sound reasons* to perform. Premise 2 makes no claims whatsoever about the psychological mechanisms that prompt actions. It says, rather, that what *makes* an act rational is that it bears the right relationship to the actor's beliefs and desires. When read this way, premise 2 has nothing to fear from empirical psychology. [10]

The EU-thesis is the most controversial premise of the DBA. As Ramsey (1931 , 174) expressed it, "I suggest that we introduce as a law of psychology that behavior is governed bymathematical expectationIf X is a proposition about which [an agent] is doubtful, any goods or bads for whose realization X is in his view a necessary and sufficient condition enter into his calculations multiplied by the same fraction, which is called the degree of his belief in X." We will discuss the general EU-thesis shortly, but the DBA requires only a special case. Assume that: (a) the agent desires *only* money; (b) her desire for money does not vary with changes in her fortune; and (c) she is not averse to risk or uncertainty. The key insight of the DBA is that, if we ignore the fact that offering someone a bet on X can alter her opinions, *the EU-thesis entails that a person satisfying (a)–(c) will reveal the strengths of her beliefs in her betting behavior*.

Suppose we offer our agent a *wager* W = [\$$a$ if X, \$$b$ else] in which she is *certain* that she will get \$$a$ if X is true and \$$b$ if X is false, and that X's truth does not depend causally on W. The EU-thesis then entails that the agent's level of confidence in X will be revealed by the monetary value she puts on W. Her *fair price* for W is that sum of money \$$f$ at which she is indifferent between receiving a payment of \$$f$ or having W go into effect. When the agent has a definite degree of belief $c(X)$ for X she will value W by its *expected payoff*, so $f = Exp(W) = c(X) \times a + (1–c(X)) \times b$. Her fair price for W is then related to her degree of belief in X by the equation $c(X) = (f–b) \div (a–b)$. [11] So, if she is indifferent between \$63.81 and a bet that pays \$100 if it rains and \$0 if not, then she is confident to degree 0.6381 that it will rain. [12]

Once this connection between degrees of belief and fair prices is granted, the DBA becomes an exercise in mathematics. Given (a)–(c), the EU-thesis entails that the agent will choose to swap any set of wagers for the sum of their fair prices, or swap any set of fair prices for its associated sequence of wagers. This "package principle" [13] entails that a person who sets prices \$0.25, \$0.25, and \$0.6 on W_X = [\$1 if X, \$0 else], W_Y = [\$1 if Y, \$0 else] and W_{XY} = [\$1 if XY, \$0 else], respectively, will exchange a portfolio containing \$0.6 and the first two wagers for one containing \$0.5 and the third. When X and Y are logically incompatible she buys nothing with her dime since the combination of W_X and W_Y is identical, in terms of payoffs, to W_{XY} . When an agent takes a self-defeating action of this sort, she is said to have made "Dutch book" against herself. Coherence entails that doing so is irrational.

The Dutch Book Theorem shows that susceptibility to Dutch books is the penalty for transgressing the laws of probability.

Dutch Book Theorem. Imagine an EU-maximizer who satisfies (a)–(c) and has a precise degree of belief for every proposition she considers. If these beliefs violate the laws of probability, then she will make Dutch Book against herself.

One proves this by showing that someone who violates a given law of probability is thereby committed to buying and selling a series of wagers whose net effect will be to cost her money come what may. For example, if one violated Additivity by having

degrees of belief 0.25, 0.25, and 0.6 for *X*, *Y*, and *XY*, respectively, then one will be susceptible to the Dutch book just described. All other violations of the laws of probability have similar "Dutch book" justifications.

The DBA has three shortcomings. It assumes an agent who meets conditions (a)–(c), who sets a fair price on every wager she considers, and who maximizes expected utility. The first two restrictions are unrealistic, and the claim that practical rationality requires EU-maximization is controversial. An adequate justification of Probabilistic Consistency must both relax the first two restrictions and provide an independent justification for EU-maximization.

Fair prices can be avoided if rational agents are permitted to have incomplete or imprecise preferences that are *extendible* to completely precise preferences that avoid Dutch books. [14] If one accepts this *requirement of coherent extendibility*, then for any rational agent there will be at least one, but usually many, complete sets of coherent fair prices that are consistent with her preferences. By the Dutch Book Theorem, each such system determines a probability function, and so the person's *Con* set will contain at least one probability.

While many Bayesians find this solution appealing, it would be better to avoid the detour through coherent extendibility altogether by showing that agents with imprecise or incomplete beliefs that violate the laws of probability are susceptible to Dutch books outright. While no general proof of this sort exists, one can go a long way toward the goal in some special cases. Suppose belief strengths can be

characterized *comparatively* by a relation $X .>. Y$ that holds when the agent is at least as confident in *X* as in *Y*. This relation need *not* be complete: neither $X .>. Y$ nor $Y .>. X$ needs to hold since the agent might have no determinate view about the relative likelihood *X* and *Y*. Even under these weak conditions, probabilistic inconsistency leaves the agent susceptible to Dutch book. [15] This result can be generalized to apply to comparative conditional beliefs and to beliefs with "interval-valued" characterizations. Hence, under broad conditions, expected utility maximizers with probabilistically inconsistent beliefs, even imprecise ones, are susceptible to Dutch books.

To relax (a)–(c) and defend the EU-thesis, we must leave the DBA behind. What is needed is a *representation theorem* that justifies probabilistic consistency and expected utility maximization simultaneously, but without restricting the *content* of the agent's desires. [16] A number of such theorems can be found in the literature, but the one due to Leonard Savage ([1954] 1972) has attained the status of a "standard model." Savage imagines an agent who uses her beliefs about the world's possible *states* and her desires for *consequences* to form *preferences* among what he calls *acts*. [17] Acts, states, and consequences are individuated so that (i) each act/state pair produces a *unique* consequence that settles every issue the agent cares about, and (ii) she is convinced that her behavior will make no causal difference to which state obtains. The agent is assumed to have a *preference ranking* over acts. For any acts *A* and *B*, she might strictly prefer *A* to *B*, be indifferent between them, or strictly prefer *B* to *A*.

In Savage's framework, an *expected utility* is a function

$$Exp_{P,u}(A) = \Sigma_{S \in States} P(S) \times u(A, S)$$

where A is an act, P is a probability over states, and u is a *utility* function. $u(A, S)$ measures the degree to which the agent's desires will be satisfied by the consequence produced by A and S. $Exp_{P,u}(A)$ is her estimate of the degree to which A is likely to produce consequences that satisfy her. Within Savage's framework, we can state the basic Bayesian requirement of practical rationality as follows:

EU-coherence. There must be at least one probability P defined on states and one utility u for consequences that *represent* the agent's preferences in the sense that, for any acts A and B, she strictly (weakly) prefers A to B only if $Exp_{P,u}(A)$ is greater than (as great as) $Exp_{P,u}(B)$.

To justify EU-coherence as the standard of practical reason, Savage imposes a system of axiomatic constraints on preferences and proves that satisfaction of these constraints guarantees the existence of the required P and u. Here are in

end p.138

formal analogues of Savage's main axioms (where A, B, C are acts, and X and Y are disjunctions of states)

Trichotomy. The agent strictly prefers A to B, strictly prefers B to A, or is indifferent between them.

Transitivity. If the agent (strictly or weakly) prefers A to B and B to C, then she also prefers A to C (see also Sorensen, chap. 14, this volume).

"Sure Thing" Principle. If A and B produce the same consequences in every state consistent with X, then the agent's preference between the two acts depends only on their consequences when X obtains (see also Dreier, chap. 9, and Sorensen, chap. 14, this volume).

Wagers. For any consequences O_1 and O_2, and any event X, there is an act [O_1 if X, O_2 else] that produces O_1 in any state that entails X and O_2 in any state that entails X.

Savage's P4. If the agent prefers [O_1 if X, O_2 else] to [O_1 if Y, O_2 else] when O_1 is more desirable than O_2, then she will also prefer [O_1 * if X, O_2 * else] to [O_1 * if Y, O_2 * else] for any other outcomes such that O_1 * is more desirable than O_2 *.

Savage showed that these axioms, along with a few others, guarantee the existence of a unique probability P and a utility u, unique up to the arbitrary choice of a unit and zero-point, whose associated expectation represents the agent's preferences. [18]

Many Bayesians use this result to justify probabilistic consistency. P4 is the lynchpin of their case. Savage uses P4 to "define" what it means for the agent to be more confident in X than in Y:

A practically rational agent *believes X more strongly than she believes Y* if and only if she strictly prefers [O_1 if X, O_2 else] to [O_1 if Y, O_2 else] for some (hence *any*, by P4) outcomes with O_1 more desirable than O_2.

Savage treated this as a "definition" because, like many social scientists of his day, he believed that legitimate objects of scientific study must be operationally defined. Strengths of beliefs are defined in terms of preferences, which are operationally defined in terms of overt choices. There are many well-known problems with this outmoded behaviorist methodology, but we need not endorse it to recognize that Savage's principle correctly captures the relationship between prefer

end p.139

ences and belief strengths (albeit not as a matter of definition). If O_1 is preferred to O_2, then the agent has *a good reason* for preferring [O_1 if X, O_2 else] to [O_1 if Y, O_2 else] exactly if she is more confident in X than in Y. Given this, P4 entails that the agent's beliefs are represented by the probability in any (P, u) pair that represents her preferences. Probabilistic Consistency is thus derived as a consequence of the theory of practical rationality embodied in Savage's axioms.

There is a vast literature dedicated to determining whether or not these axioms really are requirements of practical rationality. Savage's defenders seek to show that violating them leads agents to choose means that are necessarily insufficient for their ends. Critics try to refute these arguments, typically by arguing that they distort rational attitudes toward risk or uncertainty. The "Sure Thing" Principle has been especially controversial, but Transitivity has been questioned as well. We cannot hope to even scratch the surface of these issues here and will leave interested readers to consult the literature. [19]

There is, however, a broader worry about both the DBA and the representation theorem approaches. These arguments can show only that it is *practically* irrational to violate the laws of probability. Some critics have wondered why this should indicate anything at all about the *epistemological status* of probabilistically inconsistent beliefs. Ralph Kennedy and Charles Chihara (1979 , 30) put the point concisely: "The factors that are supposed to make it irrational to have a [probabilistically inconsistent] set of beliefsare irrelevant, epistemologically, to the truth of the propositions in question. The fact (if it is a fact) that one will be bound to lose money unless one's degrees of belief [obey the laws of probability] just isn't epistemologically relevant to the truth of those beliefs." [20]

In responding to such worries many Bayesians invoke a form of *pragmatism*. Ramsey writes that the strength of a belief "is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it" (1931, 169). De Finetti thinks it "trivial and obviousthat the degree of probability attributed by an individual to a given event is revealed by the conditions under which he would be disposed to bet on that event" ([1937] 1964 , 101). Savage, as we saw, simply *defines* the strengths of beliefs in terms of characteristic patterns of preferences for actions. Anybody who conceives of beliefs this way, as mere causes for actions, will resist the suggestion that there is any specifically "epistemological" way of evaluating them. There is nothing more to the rationality of beliefs, they will say, than their propensity to produce practically rational actions.

Other Bayesians are loath to endorse such an uncompromising pragmatism. In addition to their role in producing actions, beliefs respond to evidence, serve as premises in theoretical reasoning, and are judged in terms of the truth or falsity of their constituent propositions. Given all this, it seems a mistake to tie the rationality of beliefs too closely to that of actions or preferences. Some who feel this way adopt a "depragmatizing" strategy and reinterpret the DBA and representation theorems so as to expose the specifically *epistemic* costs of probabilistic

inconsistency. Others offer new sorts of arguments that directly bring out the epistemic shortcomings of probabilistically inconsistent beliefs.

An early "depragmatizer," Brian Skyrms, writes that "what is basic [in the DBA or representation theorems] is the consistency condition that you evaluate a betting arrangement independently of how it is described (e.g., as a bet on *XY* or as a system of bets consisting of a bet on *X* and a bet on *Y*)The cunning bettor is simply a dramatic device—the Dutch book a striking corollary—to emphasize the underlying issue of coherence" (1984a , 21–22). [21] In other words, the cognitive flaw associated with probabilistic inconsistency is a susceptibility to what psychologists call *framing effects*, in which *a single* option is evaluated differently when presented under different guises. [22] To illustrate, suppose that a person is both more confident in *X* than in *Y* and more confident in *YZ* than in *XZ* where *Z* is incompatible with *X* and *Y*. Savage's axioms entail that, for any outcomes with O_1 strictly preferred to O_2 , the agent will prefer *Package-A* = {[O_1 if *X*, O_2 else], [O_1 if *Z*, O_2 else]} to *Package-B* = {[O_1 if *Y*, O_2 else], [O_1 if *Z*, O_2 else]} but prefer *Wager-B* = [O_1 if *YZ*, O_2 else] to *Wager-A* = [O_1 if *XZ*, O_2 else]. This is inconsistent because the packages are just the corresponding wagers redescribed. While this diagnosis is correct as far as it goes, it does not reveal what is wrong with the agent's *beliefs*, since the inconsistency holds among *preferences*. [23] Some depragmatizers hope to go further by showing that inconsistent preferences are invariably accompanied by inconsistent beliefs, specifically by inconsistent beliefs about the values of prospects. [24] In an attempt to "depragmatize" the DBA, Colin Howson and Peter Urbach *define* a person's degree of belief in *X* as the betting odds she *believes* to be fair, and go on to emphasize that "believing certain odds fair does not in any way imply that you will accept bets at those odds or even at any greater odds. To believe odds to be fair is to make an intellectual judgment, notto possess a disposition to accept particular bets when they are offered" (1989, 57). Thus, someone who strictly (weakly) prefers *A* to *B* is thereby committed to the belief that it would be advantageous (fair) to have *A* rather than *B*. By the same token, a person who is more confident in *X* than in *Y* is committed to believing that Package-A will be more advantageous to her than Package-B. Given this, Howson and Urbach argue the DBA shows that probabilistically inconsistent beliefs lead to *logically* inconsistent beliefs about the values of wagers. For example, an agent who is more confident of *X* than *Y* and more confident of *YZ* than *XZ* will inconsistently believe that it is in her interest both to have Package-A over Package-B and to have Wager-B over Wager-A. [25]

Unfortunately, the Howson/Urbach approach is still infected with pragmatism. There is an inferential gap between the probabilistic inconsistency of beliefs about nonevaluative propositions and the logical inconsistency of value judgments about prospects involving these propositions. [26] When beliefs justify value judgments, the sort of "justification" at work is not purely epistemic; it invariably hinges on substantive principles of *practical* rationality. For example, the fact that the agent is more confident in *X* than in *Y* only "justifies" the judgment that Package-A is more desirable than Package-B if Savage's Sure Thing Principle is assumed. In reality, one cannot *deduce* inconsistencies in an agent's beliefs from inconsistencies in her preferences since any such inference will be mediated by the (nonlogical) principles of *practical* reasoning that relate beliefs to preferences. Whatever defects there are in the agent's beliefs are still, at root, pragmatic.

A truly epistemic rationale for Probabilistic Consistency must explain how violations of the laws of probabilities impede the *accuracy* of beliefs. Bas van Fraassen and Abner Shimony have argued, in different ways, that violating these laws can lead people to make poorly *calibrated* estimates of relative frequency. [27] Additivity requires a person who assigns a degree of belief to each proposition in a set $X = \{X_1, X_2, ,, X_n\}$ to estimate the frequency of truths in X as $[c(X_1) + c(X_2) ++ c(X_n)] \div n$. One evaluates the accuracy of such estimates using a quantity called the *calibration index*. [28] For each $1 > a > 0$, let X_a be the set of propositions to which c assigns value a, let n_a be the cardinality of X_a, and let $\alpha(\boldsymbol{X}_a)$ be the proportion of truths in $\boldsymbol{X}_a$. Then c's calibration index is given by

$Cal(c) = \Sigma_a n_a \times (\alpha(X_a) - a)^2$.

c is *well-calibrated* to the extent that it minimizes this quantity. When c is perfectly calibrated, half the propositions assigned probability 1/2 are true, two-fifths of those assigned probability 2/5 are true, and so on. Van Fraassen and Shimony each use the calibration index to measure the "fit" between beliefs and the world, and each shows that (under fairly restrictive conditions) beliefs that violate the laws of probability are necessarily less well calibrated than they could otherwise be. These are interesting results, but they do not justify probabilistic consistency because *calibration simply is not a reasonable measure of the "fit" between graded beliefs and the world*. To state just one problem, Joe can be better calibrated than Jane even though Jane *always* believes truths more strongly, and falsehoods less strongly, than Joe does. [29]

Another approach, pursued in Joyce 1998 , relates probabilistic consistency directly to the *accuracy* of graded beliefs. The strategy here involves laying down a set of axiomatic constraints that any reasonable gauge of accuracy for confidence measures should satisfy, and then showing that probabilistically inconsistent measures are always less accurate than they need to be. The two most important constraints are

Normality. If c's values are always at least as close to the actual truth-values of propositions as c^*'s are, then c is at least as accurate as c^*.

Convexity. If c and c^* are equally accurate (and not identical), then their *even mixture* $[c + c^*] \div 2$ is strictly more accurate than either.

end p.142

Normality connects accuracy to truth. *Convexity* penalizes overconfidence (so that, e.g., believing X and X both to degree 1 2 is better than believing them both to degree 1). On the basis of these axioms, and a few others, it can be proved that if c violates the laws of probability then there is a probability function c^+ that is strictly more accurate than *cunder every logically consistent assignment of truth-values to propositions*. To the extent that one accepts the axioms, [30] this shows that the demand for probabilistic consistency follows from the purely epistemic requirement to hold beliefs that accurately represent the world.

3. Bayesian Confirmation Theory

The most influential aspect of Bayesian epistemology is its theory of *evidential support*. Bayesians reject the idea that evidential relations can be characterized in an objective, belief-independent manner; evidence is always relativized to a person and her opinions. On this view, a person's *total*, *nonincremental* evidence regarding a hypothesis *H* is directly reflected in her level of confidence in *H*. [31] This evidence derives from two sources: (a) the person's own subjective "prior" opinions about the intuitive plausibility of *H* and other propositions, and (b) any new knowledge she has acquired via learning. She is more confident in *H* than in *H* exactly if the totality of her prior and learned evidence tells more strongly in favor of *H* than against it. Similarly, her level of confidence in *H* conditional on *E* reflects her *total* evidence for *H* when *E* added to her stock of knowledge. The *disparity* between her unconditional level of confidence in *H* and her level of confidence in *H* given *E* then captures the amount of *additional* evidence that *E* provides for *H*. This justifies the following qualitative account of *incremental* evidential support:

Econfirms(disconfirms, is *irrelevant* to) *H* for a person whose beliefs are represented by the measures in *Con* if and only if *c*(*H*/*E*) exceeds (is exceeded by, equal to) *c*(*H*) for *everycCon*.

Bayesians use this analysis to provide explanations of important truisms about evidential relationships and facts about scientific practice. [32] Here are two examples that both follow from Bayes's Theorem:

Prediction Principle. If a person is more confident in *E* conditional on *H* than conditional on *H*, then *E* confirms *H* for her.

end p.143

Surprise Principle. If a person is equally confident in *E* and *E** conditional on *H*, then *E* confirms *H* more strongly for her than *E** does (or disconfirms it less strongly) if and only if she is *less* confident of *E* than of *E**.

The prediction principle provides a Bayesian rationale for the *hypothetico-deductive model of confirmation*. On the H-D model, hypotheses are incrementally confirmed by any evidence statements they *entail*. Bayesians are able to make sense of this insight even though they reject the idea that evidential relations can be characterized independent of belief. When *H* entails *E*, the Prediction Principle entails that *E* (incrementally) confirms *H* for anyone *who does not already reject H or accept E*. Consequently, the results of experiments that fit the H-D paradigm will be deemed evidentially relevant by *anyone* who has not yet made up his mind about the data or the hypothesis. While the *degree* of confirmation will vary across people, every rational person will agree that the hypothesis is supported by the data to some degree.

The surprise principle explains why *unexpected* evidence seems to have more "confirming potential" than evidence that is already known. On the Bayesian view, a person's appraisal of *E*'s evidential import for *H* varies *inversely* with her confidence in *E* when *c*(*E*/*H*) is held fixed. In particular, if *H* entails *E*, so that *c*(*E*/*H*) = 1, then *E* confirms *H* more strongly the less likely it is.

Some see this as a double-edged sword. According to Clark Glymour (1980 , chap. 3), it entails that "old" evidence about which a person is already certain (or very highly confident) cannot provide any (much) support for any hypothesis. When *c*(*E*) is close to

1, C(H) will be close to $c(H/E)$, and this detracts from E's confirming power. As Glymour emphasizes, however, many splendid pieces of scientific reasoning involve adducing "old" evidence to support novel hypotheses. Many aspects of this *Problem of Old Evidence* pose challenges for Bayesianism, but the most serious is to explain how a person who is certain or almost certain of E, and is fully aware of all the relevant logical relationships between H and E, can see E as evidence for H. [33]

Such an explanation can be given once we realize that the Bayesian tent has room for more than one notion of evidential support. The point is easiest to see when put in quantitative terms. Many Bayesians use the *difference measured*$(H, E) = c(H/E)–c(H)$ to capture the *degree* to which E incrementally supports H. Since this goes to zero as $c(E)$ approaches one the Problem of Old Evidence clearly arises for d. There is, however, a closely related function, $d^*(H, E) = c(H/E)–c(H/E)$, that lacks this undesirable feature: as long as $c(H/E)$ and $c(H/E)$ are defined, $c(E)$ can range from zero to one without affecting the value of $d^*(H, E)$. The suggestion, made in Joyce 1999 and Christensen 1999 , is that d^* captures the sense of confirmation at issue in old evidence cases. It compares total evidence that a person will have for H if E is added to her stock of knowledge with the total evidence she will have for H if E is added. Since this comparison does *not* depend

end p.144

on her level of confidence in E, it can remain fixed even as her beliefs about E and H change. Thus, a person who is certain of E, and who is fully aware of all the relevant logical relationships between H and E, can still see E as evidence for H in the sense of regarding H as more likely if E is true than if E is false.

It must be emphasized that d^* is *not* being proposed as a replacement for d. [34] Rather, d remains the basic measure of incremental confirmation, and d^* captures one of its *components*. Writing $d(H, E) = c(E) \times d^*(H, E)$ makes it clear that d^* is a part of d that does *not* depend on E's prior probability. There are many questions about confirmation that d can be used to answer, but d^* is useful for answering others. We should use d when we want to say how much E supports *Hrelative to a common background of knowledge concerning E*. d^* is useful when we are talking about E's confirming power *across* states of knowledge about E in which $c(H/E)$ and $c(H/E)$ are fixed. (This turns out to be a fairly common case.) Hence, there is no *conflict* between d and d^*: they measure different things and are used to answer different questions.

There are other measures of confirmation as well. The most important is the *likelihood ratioc*$(E/H) \div c(E/H)$, which expresses the disparity between $c(H/E)$ and $c(H)$ as the ratio of H's *odds* given E to its unconditional *odds*. [35] This measure is useful when there is consensus about how strongly H and *HpredictE*, but none about the probabilities of H and E. Here too it is best to be ecumenical. There is no single right way to measure confirmation; different measures are suited for different pragmatic purposes. To tell the *whole* story about a person's evidential situation vis-à-vis H and E, we would need to describe her *entire* sys-tem of beliefs. Fortunately, the whole story rarely interests us. We use different senses and measures of incremental confirmation to illuminate those parts of it that do.

That said, it is possible to isolate an element that is common to any theory of evidence that deserves the name *Bayesian*. The core tenet of any such theory is the observation,

encapsulated in Bayes's Theorem and the Prediction Principle, that *E's confirming potential relative to H is enhanced to the extent that H makes E more probable*. It follows from this that the ability of evidence to *discriminate* among competing hypotheses is limited by the relative degree to which these hypotheses probabilify the evidence. Imagine a believer who has more *total* evidence for *H* than for *H**, so that $c(H) > c(H^*)$. What would it take to reverse this situation by adding *E* to her knowledge, so that $c(H^*/E) > c(H/E)$? Bayes's Theorem tells us that *E* can reverse the "balance of evidence" only if *H** predicts *E*'s truth more strongly than *H*'s does. This leads to the following: *Discrimination Principle*. If a person initially has more total evidence for *H* than for *H**, and if *H* predicts *E* at least as strongly as *H** does, then the person will have more total evidence for *H* than for *H** after *E* is added to her stock of knowledge.
end p.145

Note that *E* can *never* reverse the balance of total evidence between *H* and *H** when both hypotheses *entail E*. This observation will play a central role in our discussion of the Bayesians' account of learning.
Like the rest of Bayesian confirmation theory, the Discrimination Principle rests on the assumption that a person's total evidence is a combination of her "prior" views about the intuitive plausibility of hypotheses and information she has acquired via learning. So, to fully understand Bayesianism, we need some appreciation of its conception of "prior" opinions and its theory of learning. We begin with the latter.

4. The Bayesian Theory of Learning

Bayesians see learning as a process of *belief revision* in which "prior" beliefs are replaced by "posterior" beliefs that incorporate new information. For ease of exposition we will assume an ideal learner whose prior and posterior opinions can be represented by probability functions c_0 and c_1. Bayesian learning proceeds in two stages: one causal, one inferential. First, the person has a *learning experience* in which perception, intuition, memory, or some other *noninferential causal process* immediately alters some subset of her beliefs. Second, she uses information acquired via this experience, in conjunction with other things she knows, to revise the rest of her opinions. We will focus on experiences whose sole *immediate* effect is to alter the person's level of confidence in a proposition *E*, [36] so that c_1 is constrained to be such that $c_1(E) \neq c_0(E)$. Every other change in the learner's posterior beliefs is due to a (probabilistic) inference from this basic one. The challenge is to explain how her *prior* opinions can justify the choice of a specific posterior c_1 from among the many that might meet the constraint.
In the simplest learning experiences, where the person becomes *certain* of *E*, Bayesians advocate the following rule of belief revision:
Simple Conditioning. If a person with a "prior" such that $0 < c_0(E) < 1$ undergoes a learning experience whose only immediate effect is to raise *E*'s probability to one, then $c_1(H) = c_0(H/E)$ for any proposition *H*.

The effect of this is to set probabilities of hypotheses inconsistent with E to zero and to increase probabilities of hypotheses that entail E uniformly by a factor of $1/c_0(E)$.

Simple conditioning requires a learner to become *certain* of E's truth. As Richard Jeffrey (1983 , 164–69) has long argued, however, our evidence is typically too vague and imprecise to justify such "dogmatism." More realistic learning experiences are modeled by the following rule:

Jeffrey Conditioning. If a person with a "prior" such that $0 < c_0(E) < 1$ undergoes a learning experience whose only immediate effect is to set $c_1(E) = q$, then $c_1(H) = q \times c_0(H/E) + (1–q) \times c_0(H/E)$ for any proposition H.

This holds probabilities conditional on E and E fixed, so that $c_1(H/E) = c_0(H/E)$ and $c_1(H/E) = c_0(H/E)$, and it multiplies the probability of each hypothesis that entails E (or E) by a factor of $q/c_0(E)$ (or $(1–q)/c_0(E)$). Note that Jeffrey Conditioning reduces to Simple Conditioning when $q = 1$.

Many justifications have been offered for the conditioning rules. The most contentious are the *diachronic Dutch book arguments*. Like their synchronic counterparts, these are *pragmatic* arguments designed to show that failing to condition on one's evidence can lead to practical incoherence. Here, though, the self-defeating choices are made at *different* times. David Lewis offered a Dutch book rationale for Simple Conditioning, Brad Armendt generalized it to cover Jeffrey Conditioning, and Bas van Fraassen and Michael Goldstein each showed that both forms of conditioning are instances of the *Reflection Principle*, which also has a Dutch book rationale. [37] Our discussion will focus on Reflection since it is the most general principle of the bunch.

Reflection requires a person's degree of belief in H at time t_0 to agree with her t_0-*expectation* of her time t_1 degree of belief in H. If "[su1]$c_1(H) = x$" says the person's degree of belief in H at time t_1 is x, then Reflection requires that $c_0(H/[\text{su1}]c_1(H) = x) = x$ and therefore that

$$c_0(H) = \Sigma_x\, c_0([\text{su1}]c_1(H) = x) \times x.$$ [38]

This captures what is at issue in diachronic Dutch Book arguments, since, as van Fraassen and Goldstein each showed, a person is invulnerable to diachronic Dutch book *if and only if* she satisfies Reflection.

As many authors have recognized, [39] however, there is nothing irrational, *per se*, about violating Reflection or leaving oneself open to diachronic Dutch books. In fact, practical rationality *requires* it when one undergoes an *antilearning experience*, that is, a belief change one takes to be *un*reliable. Suppose a miserly expected utility maximizer has inadvertently ingested a drug that will soon make her highly confident of the claim, H, that she will be able to recite the Gettysburg Address backward on command. She now regards H as unlikely, and does not take the fact that she will soon believe it to be any indication of its truth. What should she do? Since she knows that she will soon pay close to \$1 to buy the wager [\$1 if H, \$0 else], the best way for her to offset her future idiocy is to hedge her bets now by selling the wager for as much as she can get above its current

(low) fair price. Even though she is sure to lose money in the aggregate, this is perfectly rational. The irrational thing would be to allow the beliefs of her future, idiot self to guide her present actions.

It is irrational to violate Reflection *only* when one regards the belief change one is about to undergo as a genuine *learning experience* that is likely to improve the overall accuracy of one's beliefs. But, what is it to *regard* a belief change as a learning experience? As Brian Skyrms (1993) has argued, Reflection itself provides the answer. A person sees a prospective change in her belief about H as increasing accuracy exactly if she sees her future degree of belief in H as a reliable indicator of H's truth-value. When the evidential connection is perfect, her beliefs will satisfy $c_0(H/[su1]c_1(H) = x) = x$ for all x. So, *by definition*, a person regards a belief change as a genuine learning experience only if it satisfies Reflection. This makes Reflection (or invulnerability to diachronic Dutch books) entirely useless as a *justification* for the conditioning rules. Indeed, when *properly* formulated (as here), these rules take the shift from $c_0(E)$ to $c_1(E)$ *at face value as a learning experience*.

A further problem with diachronic Dutch book arguments is their *pragmatic* character. Even if conditioning *pays*, antipragmatists will still wonder whether it leaves learners better off from a purely epistemic perspective. One can argue that it does by showing that conditioning rules are "epistemically conservative" in that they produce a posterior that *departs minimally* from the prior while taking the evidence into account. The best results along these lines are due to Persi Diaconis and Sandy Zabell (1982), who show that, on a variety of ways of measuring differences between probabilities, simple and Jeffrey conditioning *uniquely* minimize change subject to the constraints imposed by experience. A more general approach, still grounded in epistemic conservatism, justifies the conditioning rules by showing that they alone keep central features of the prior intact. Consider the property

Ordinal Invariance. If $c_1(E) > 0$ then, for any hypotheses H and H^*, $c_1(H \,\&\, E) > c_1(H^* \,\&\, E)$ iff $c_0(H \,\&\, E) > c_0(H^* \,\&\, E)$.

This requires a person who has a learning experience involving E to retain her views about the *comparative* probability of propositions that entail E. It can be shown that *only* the conditioning rules generally meet this condition. [40] Thus, we will have a sound epistemic rationale for conditioning if we can make a convincing case for Ordinal Invariance.

The Discrimination Principle of section 3 gives us the resources to do so. It entails that a person who initially has more total evidence for $H \,\&\, E$ than for $H^* \,\&\, E$ will still have more total evidence for $H \,\&\, E$ after undergoing a learning experience that alters E's probability. As already noted, a rational believer will invest more prior (posterior) confidence in $H \,\&\, E$ than in $H^* \,\&\, E$ just in case her total evidence for the former at t_0 (at t_1) exceeds her total evidence for the

end p.148

latter. It follows directly from Discrimination that c_1 will rank $H \,\&\, E$ above $H^* \,\&\, E$ if and only if c_0 does. Hence, Discrimination implies Ordinal Invariance. Simple and Jeffrey conditioning are thus justified as the only belief revision rules that are consistent with this most basic tenet of the Bayesian theory of evidence.

5. Prior Probabilities: The Charge of Subjectivism

The most persistent objection to Bayesianism is that it engenders an untenable *subjectivism* in which all manner of preposterous beliefs and ludicrous inferences are

immunized from criticism. If the constraints on rational opinion begin and end with the laws of probability and the conditioning rules, then Bayesianism allows a person to draw almost any conclusion on the basis of almost any evidence as long as she starts out with suitable prior beliefs. There are, for example, probabilistically consistent priors that make the existence of statues on Easter Island evidence for the conclusion that Martians built the Pyramids, or that count peyote-induced belief changes as learning experiences. By allowing such absurdities, Bayesianism seems to say that what it makes sense to believe, and what counts as evidence for what, is "all just a matter of opinion." In Bayesian epistemology, it appears, just about anything goes.

Bayesians have addressed this charge in a variety of ways. "Personalists," like Savage and de Finetti, bite the bullet and deny that there are any constraints on beliefs that outrun the laws of probability and the conditioning rules. They seek to blunt the force of the "anything goes" objection by arguing that the effect of a person's priors will tend to diminish as she acquires increasingly more evidence about the world. Here is a famous statement of the view: "If observations are precisethen the form and properties of the prior distribution have negligible influence on the posterior distribution. From a practical point of view, then, the untrammeled subjectivity of opinionceases to apply as soon as much data becomes available. More generally, two people with widely divergent prior opinions but reasonably open minds will be forced into arbitrarily close agreement about future observations by a sufficient amount of data." [41] This "merger of opinion" is what stands in for objectivity in the personalist picture: objectivity is *intersubjective agreement* in the long run.

The theoretical basis of such claims is found in a set of mathematical results, the "washing out" theorems, which show that people who begin with different priors will tend to reach consensus as they acquire increasingly more data. Sup

end p.149

pose we have two subjects with priors c and c^* that assign H an intermediate probability. Imagine further that there is an infinite sequence of evidence statements $\{E_1, E_2, E_3, \}$ such that for each j:

(a) c and c^* assign each finite data sequence $D_j = \pm E_1 \,\&\, \pm E_2 \,\&\&\, \pm E_j$ an intermediate probability. [42]

(b) At time-j each subject has an experience that sets E_j's probability to 1 or to 0.

(c) Both subjects regard these as *learning* experiences.

(d) Both subjects condition on the evidence they receive, so that for each j and each finite data sequence D_j, $c_j(H) = c(H/D_j)$ and $c_j{}^*(H) = c^*(H/D_j)$.

(*a*)–(*d*) entail that the probabilities each subject assigns to H at successive times form a *martingale sequence* in which each term is the expected value of its successor, so that c_j

$(H) = \Sigma_x c_j (c_{j+1}(H) = x) \times x$. The *Doob Martingale Convergence Theorem* then entails that, with probability one, $c_j(H)$ and $c_j^*(H)$ each converge to a definite limit. [43] Establishing that these limits coincide requires an additional assumption. There are two kinds of conditions that will do the job. First, the evidence can be so informative that it forces *any* rational believer to the same conclusion about *H*. Either of the following clauses will accomplish this:

(e) Each possible data sequence $\pm E_1 \,\&\, \pm E_2 \,\&\, \pm E_3$,entails *H* or $\overline{H}$.
(*e**) Each data sequence determines an *objective probability* for *H*.

On the basis of Coherence alone, *(e)* guarantees that $c_j(H)$ and $c_j^*(H)$ will converge to *H*'s truth-value (as specified by the data). Similarly, if we combine *(e*)* with what David Lewis (1980) calls the "Principal Principle," which requires rational believers to align their degrees of belief with known objective probabilities, [44] then $c_j(H)$ and $c_j^*(H)$ will converge to *H*'s objective probability (as specified by the data).
For purposes of allaying fears about subjectivism, these ways of obtaining agreement in the limit are wholly impotent. Priors only get "washed out" because the data is so incredibly informative that, as a matter or logic, it makes each subject's antecedent beliefs *irrelevant* to her final view. Indeed, a "washing out" result that uses (*e*) (or [*e**]) amounts to little more than the claim that two subjects will come to agree about *H*'s truth-value (or probability) if each ultimately learns *H*'s truth-value (or its objective probability). If all learning situations were like this, there would be no need for prior probabilities at all.
To obtain washing-out theorems that respond to subjectivist worries, we must imagine that nothing in the laws of logic or probability forces subjects to assign a specific probability to *H* conditional on each data sequence. We must, rather,
end p.150

derive joint convergence from commonalities among the priors alone. One way to do this, pioneered by Savage ([1954] 1972 , 46–50), is to have subjects agree that the evidence statements are *statistically independent and identically distributed* conditional on *H* and $\overline{H}$.

(*e***)For some constants *x* and *y* and all *j* and *k*, $c(E_k/H) = c(E_k/H \,\&\, E_j) = x$ and $c(E_k/\overline{H}) = c(E_k/\overline{H} \,\&\, E_j) = y$. The same holds with *c* replaced by *c**.

Savage showed that under these conditions $c_j(H)$ and $c_j^*(H)$ converge to the same value with probability one (according to both *c* and *c**).
There is less here than meets the eye. (*e***) requires an *exceptional* amount of initial consensus between the parties. In effect, both must start out treating *H* like a chance hypotheses, so that *H* = "*x* is the objective probability of every E_j ," and each must see the E_j as describing independent trials of a chance process. While this does happen in

rare, well-behaved cases (e.g., coin flipping), it typically fails. Once again, the assumptions needed to ensure convergence to a common value severely limit the theorem's value as a response to the charge of subjectivism.

Indeed, it is hard to see how any result based on *(a)–(d)* can rebut subjectivism. These assumptions require *substantial* initial agreement among the subjects. *(a)* asks them to agree about which data sequences have a chance of occurring. *(c)* requires them to see the belief changes they are about to undergo as learning experiences. If they disagree about these things, they will *not* tend toward consensus as the evidence accumulates simply because they will not agree about what counts as "the evidence." The basic problem with using "washing out" results to refute subjectivism is simple: no agreement in, no agreement out!

Another approach, championed by so-called *objective* Bayesians, seeks to avoid subjectivism by restricting prior probabilities. The best-developed view of this sort is that of the physicist E. T. Jaynes, who writes, "The most elementary requirement of consistency demands that two persons with the same relevant prior information should assign the same prior probabilities. Personalistic doctrine makes no attempt to meet this requirement." [45] According to Jaynes, any well-posed problem of inductive inference is defined by certain *objective constraints*, often deducible from physical theory or symmetry principles, which fix *expected values* for various quantities. An *acceptable* prior must yield the required expectations. If the expected value of the toss of a die is constrained to be 3.5, say, then $c(1) \times 1 + c(2) \times 2 + \ldots + c(1) \times 6 = 3.5$ will hold for every acceptable prior c. To choose the correct prior from among the many that are acceptable, Jaynes advocates *entropy maximization*. Relative to a partition $\{X_1, X_2, \ldots, X_k\}$ of mutually exclusive, collectively exhaustive propositions, the entropy in a probability c is defined as

$$\textit{Entropy}(c) = \Sigma_j \, c(X_j) \times \log(c(X_j)).$$

Entropy(c) measures the inverse of the amount of *information* (about the X_j) that c encapsulates. For the die above it is easy to show that the uniform distribution, in which each side comes up with probability 1/6, uniquely maximizes entropy. By choosing the (unique) probability that maximizes entropy, Jaynes argues, we respect the constraints but otherwise make the fewest possible additional assumptions.

While maximizing entropy is a fine method for finding a probability that meets specific constraints, the idea that it "objectifies" Bayesianism is illusory. As Colin Howson and Peter Urbach (1989 , 289) point out, the choice of one system of constraints rather than another is a subjective matter par excellence: "No prior probabilityexpresses merely the available factual data; it inevitably expresses some sort of opinion about the possibilities consistent with the data. Even a uniform probability distribution is [uniform] only relative to some partition of these possibilities: we can always find another with respect to which the distribution is as biased as you like—or don't like. Jaynes's objective priors do not exist." Even if we all agree that the expected value of a toss of a die is 3.5, we will be justified in settling on the uniform distribution only if we also agree that this is our *only* relevant piece of information. This is a *subjective* judgment, as is any other of the form "precisely *these* constraints characterize our knowledge." If you think that odd tosses are more likely than even ones, and I think the reverse, then we would both be wrong to ignore our beliefs and settle on the uniform distribution. We might agree with Jaynes that

"consistency demands that two persons with the same relevant prior information should assign the same prior probabilities," but we will disagree about which "prior information" counts as relevant and even about what this information is. Appealing to a physical theory will not help matters unless we are already convinced of it, which, again, is a matter for our respective background beliefs. Jaynes's program provides no answer to the charge of subjectivism.

Perhaps Bayesians should admit that there is more to epistemic rationality than the laws of probability and the conditioning rules. These provide an *internalist logic* of rational belief: they tell us whether a person's beliefs cohere with one another, what she counts as evidence for what, and how she should revise her opinions in light of what she regards as learning experiences. The whole Bayesian apparatus, in other words, is appropriate for describing and criticizing a person's *own reasons* for believing what she does.

Any internalist view is going to suffer from "garbage in, garbage out" problems. If a person starts out with inaccurate priors that assign low probabilities to truths and high ones to falsehoods, or if she has false views about which belief changes are learning experiences, then her subsequent beliefs will be inaccurate and unreliable as well. Sophisticated Bayesians should concede this and grant that there are further *externalist* principles of epistemic rationality that can be used to evaluate opinions. These principles will not assess a person's beliefs on the basis

of *her own* view of things—that is, they will not take her prior opinions or her views about learning at face value as the personalists do. Rather, they will consider the actual accuracy of her priors and the actual reliability of the processes she treats as learning experiences. Ramsey saw this right from the start. In addition to being concerned with *rational* belief, which answers to internalist Bayesian norms, he sought a theory of *reasonable* belief that would assess a person's doxastic attitudes and habits on the basis of the actual accuracy of the beliefs they generate. [46] Personalist Bayesians have largely ignored this aspect of Ramsey's thought, but they would be wise to give it more heed.

One should not think of these externalist approaches *replacing* Bayesianism. A complete account of epistemic rationality will necessarily involve both internalist considerations that concern a believer's reasons and externalist considerations having to do with the accuracy of her opinions and the reliability of her belief-forming processes. While Bayesians have little to say about the latter issues, they offer us a detailed, systematic, and exceedingly plausible account of the former. For all its shortcomings, *Bayesianism remains without peer as a theory of epistemic reasons and reasoning*. As long as we use it for this purpose it will serve us well.

NOTES

1. Classic sources are Ramsey 1931 , de Finetti (1937) 1964 , and Savage (1954) 1972 .

2. The strength of a belief should not be confused with any *feeling of conviction*. As Ramsey (1932, 169) noted, the beliefs we hold most strongly are often associated with no feelings whatever. Moreover, people who *feel* convinced of propositions sometimes

reason and act as if they are false. Nor is a graded belief a categorical belief about an *objective probability*. A person can be *certain* that the coin he is about to toss is either two-headed or two-tailed, and yet be maximally uncertain about which possibility obtains. He then believes "the coin will come up heads" to degree 1/2 even though he knows that 1/2 is *not* its objective probability. In general, a person's degree of confidence in a proposition is her *subjective expectation* of its objective probability.
3. See Rényi 1955, Harper 1976 , Spohn 1986 , McGee 1994 , van Fraassen 1995 , and Hammond 1994 .
4. See Levi 1980 , 85–91, and Kaplan 1996 , 27–31.
5. *P* must also be *countably* additive. The issues surrounding countable additivity are too involved to pursue here. See Seidenfeld and Schervish 1983 and Kaplan 1996 , 32–36.
6. Conditioning entails that $P(X/Y) = P(X \,\&\, Y) \div P(Y)$ when $P(Y) > 0$.
7. Bayes (1764) 1958 .
8. See Osherson 1995 and Shafir and Tversky 1995 .
9. Ramsey is especially lucid on this point in 1931, 173.
10. It need not be any part of Bayesianism that acting on the basis of habits, emo

tions, and so on is irrational. If an act bears the right relationship to the actor's beliefs and desires, then it is rational however it is caused. When emotions or habits tend to lead (nonaccidentally) to rational actions, Bayesians should *encourage* emotional or habitual decision making.
11. The value of $c(X)$ is independent of the particular choice of a and b. For any $a^* > b^*$, the fair price f^* of $W^* = [\$a^* \text{ if } X, \$b^* \text{ else}]$ will be such that $c(X) = (f^*–b^*) \div (a^*–b^*)$.
12. Conditional beliefs are reflected in fair prices of bets that get "called off" when the condition fail to obtain.
13. Some commentators incorrectly portray the "package principle" as an added, hidden premise in the DBA. Actually, it follows from the EU-thesis and (a)–(c).
14. See Kaplan 1996 , 23–31; Jeffrey 1992 , 82–85; Joyce 1999 , 43–45.
15. The proof of this claim, which is beyond the scope of this paper, relies on technical results found in Kraft, Pratt, and Seidenberg 1959 and Scott 1964 .
16. Most Bayesians regard the DBA as a "toy model" whose rhetorical purpose is to point beyond itself to these more serious representation results. See Skyrms 1984b , Jeffrey 1992 , Maher 1993 , Kaplan 1996 , and Joyce 1999 , among others. This has *always* been the prevailing wisdom among Bayesians. Indeed, Ramsey only mentions the Dutch Book Theorem in passing *after* having proved a representation theorem.
17. Despite this terminology, Savage's acts are *not* best understood as events the agent can directly control. See Joyce 1999 , 61–62, and 107.
18. To establish uniqueness Savage imposed a variety of constraints (like Trichotomy and Wagers) that require extremely *rich* systems of preferences. Most Bayesians now agree that these *richness conditions* far exceed what is demanded by practical rationality. Accordingly, Savage's result is best thought of as guaranteeing the existence of a (large) family of probability/utility pairs, all of whose associated expectations represent the agent's preferences.
19. See Broome 1991 and Maher 1993 for overviews.

20. For similar worries, see Rosenkrantz 1981 , 214; Joyce 1998 , 584–86; and Christensen 2001 , 356–64.
21. Skyrms claims, rightly, that Ramsey too saw this as the deep flaw that incoherent preferences serve to indicate. Armendt (1993 , 3) defends a similar view.
22. See Tversky and Kahneman 1981 .
23. See Joyce 1998 , 586.
24. See Howson and Urbach 1989 , Christensen 1996 , Hellman 1997 .
25. Christensen (2001) seeks to reinterpret representation theorems in an analogous way.
26. Compare Maher 1997 .
27. See van Fraassen 1983 and Shimony 1988 .
28. For useful discussion of the calibration index, see Murphy 1973 .
29. See Joyce 1998 , 494–95. For other problems, see Seidenfeld 1985 .
30. Maher (2002) expresses reservations about *Convexity*, and champions a nonconvex accuracy gauge.
31. While Bayesians rarely discuss *total* evidence, their theory of incremental evidence makes sense only when it is based on an account of total evidence like the one given here.
32. For useful discussion, see Earman 1992 , chap. 3.
end p.154

33. Other aspects of the problem concern how to characterize the knowledge of logical relationships within the Bayesian framework, and how to handle genuinely novel hypotheses. See Joyce 1999 , 204, and Earman 1992 , 120–35, for relevant discussion.
34. This runs contrary to the line taken in Eells 2000 .
35. *H*'s *odds* are $c(H) \div c(H)$, and its *odds conditional on E* are $c(H/E) \div c(H/E)$. Odds talk can be translated into probability talk using the mapping $c(H) = odds(H) \div (1 + odds(H))$.
36. More complicated experiences might alter her levels of confidence in each element of a set of statements, or alter the conditional probability of some proposition given another, or directly fix the value of a random variable.
37. Lewis's result is reported and discussed in Teller 1973 . Armendt 1980 , van Fraassen 1984 , and Goldstein 1983 contain the generalizations.
38. As van Fraassen notes, it is important to understand that [su1]$c_1(H)$ is a *nonrigid* designator since the identity [su1]$c_1(H) = x$ is otherwise trivial.
39. Levi 1987 ; Christensen 1991 ; Maher 1993 , 106–20.
40. Joyce 1999 , 195–96.
41. Edwards, Lindeman, and Savage, 1963 , 201. Also, see Suppes 1966 , 204.
42. $\pm E$ may be E or $\overline{E}$.
43. Doob 1971 .
44. Of course, personalists, like de Finetti and Savage, would deny that objective probabilities exist.
45. Jaynes 1968 , 53. See also Jaynes 1994 .
46. See Ramsey 1931 , 193–96.
end p.155

Chapter 9
DECISION THEORY AND MORALITY
James Dreier

This chapter is about decision theory (see also Joyce, chap. 8, this volume) and morality. Its main point is to show how the formal apparatus of decision theory is connected to some abstract issues in moral theory. It does not aim to be comprehensive.
The preliminaries section explains how to think about utility and the advice that decision theory gives us. In particular, decision theory does *not* assume or insist that all rational agents act in their own self-interest. The second section discusses decision theory's contributions to social contract theory, with emphasis on David Gauthier's rationalist contractualism. The third section considers a reinterpretation of the formal theory that decision theorists use: utility might represent goodness rather than preference. The last section discusses Harsanyi's theorem.

1. Preliminaries

According to decision theory, rational agents always choose the alternative with the greatest utility, which is also the one with the greatest *expected utility*. Suppose Liz is interested in the expected utility of applying to two graduate programs in philosophy, one at Prestige U. and the other at Humble State. She can afford only one application. There are three possible final outcomes: Liz may be admitted to Prestige, or she may be admitted to Humble, or she may be admitted nowhere. Let's call these outcomes P, H, and N. Each application may be thought of as a *lottery*. [1] There is some chance, say p, that Liz will be admitted to Prestige if she applies, and there is some chance, say q, that she will be admitted to Humble State if she applies there. We can use the notation "$L[p,x,y]$"
for a lottery that yields x as prize with probability p, and otherwise (with probability $1–p$) yields prize y. So Liz's choices are $L[p, P, N]$ and $L[q, H, N]$. If we use "$u(x)$" to denote the utility of x, then the formula for calculating the expected utility of a lottery in general is
$eu(L[p,x,y]) = p\times u(x) + (1–p)\times u(y)$
So for Liz the expected utilities for the two applications are
$eu(L[p,P,N]) = p\times u(P) + (1–p)\times u(N)$, and
$eu(L[q,H,N]) = q\times u(H) + (1–q)\times u(N)$
Decision theory says that if Liz is rational, she will perform the act whose expected utility is higher. She can calculate the expected utility of applying if she can fill in the terms in the formula with numbers. We will assume that our agents can make determinate probability judgments about all sorts of things. Then Liz need only decide what utility numbers to assign to the possible outcomes. Easy? Difficult? Impossible? How should she begin?
What Utility Is
It is sometimes said that utility is supposed to be a measure of the agent's own self-interest. If that were so, the theory would be telling each agent to maximize the

expectation of her own self-interest. And in that case, it would conflict dramatically with the demands of morality. But standard decision theory does *not* use any measure of an agent's self-interest as a utility function (see also Weirich, chap. 20, this volume). Even so, there is a sense in which decision theory may make demands on an agent that conflict with the demands of morality. The standard interpretation takes utility to be a measure of the agent's preferences.
end p.157

The option with the highest utility number for Liz is the option she most prefers. Only if Liz is unusually selfish will her utility function represent her own self-interest. On the other hand, unless Liz always prefers the morally required option, her preferences will sometimes conflict with the demands of morality. Understood as a theory of rationality, [2] decision theory says that it may sometimes be rational to act contrary to one's moral requirements. At least, it says so on the assumption that moral requirements are what they seem to common sense to be.

Now Liz knows that she is supposed to assign higher numbers to the preferred outcomes, but how much higher? Liz needs more advice from decision theory about how to proceed. The standard advice given involves asking Liz some questions about her preference ranking, especially about how she ranks various artificially constructed lotteries. Rather than go through this standard advice, I will explain the central theorem of decision theory and the axioms from which the central theorem is derived (but I will not prove the theorem). The theorem guarantees that preferences that satisfy the axioms will also be representable by an expectational utility function.

The Axioms

There are many ways of presenting the axioms of decision theory. I will use Michael Resnik's version of the axioms. [3]

The axioms concern a relation, *R*, thought of as representing a person's preferences. Expressions of the form *xRy* mean that the agent in question either prefers *x* to *y* or is indifferent between them.

First, we want the relation to order the field of alternatives (allowing ties). That means it must meet the following constraints:

(O1)(*x*)(*y*) (*xRy* ν *yRx*) [That is, the ordering is *complete*.]

(O2)(*x*) (*xRx*) [The relation is *reflexive*.]

(O3)(*x*)(*y*)(*z*) [(*xRy* ˜ *yRz*) → *xRz*] [The relation is *transitive* (see also Joyce, chap. 8, and Sorensen, chap. 14, this volume).]

It is helpful to have an abbreviation for "*xRy* ˜ y*R*x," so for that we'll use "*x*~*y*"; intuitively this means that the agent is indifferent between *x* and *y*.

Next, we need some axioms that concern what we called "lotteries." The first lottery axiom is called *continuity*.

(L1) $(x)(y)(z)\{(xRy \tilde{} yRz) \rightarrow p(y \sim L[p,x,z])\}$

The idea is that if you prefer x to y and y to z, then there must be some lottery between x and z such that you would be indifferent between that lottery and y.
The second lottery axiom is called *better prizes*.
end p.158

(L2) $(x)(y)(z)(p)(xRy \leftrightarrow L[p,z,x]\ RL[p,z,y])$ and $(x)(y)(z)(p)(xRy \leftrightarrow L[p,x,z]\ RL[p,y,z])$

Intuitively, if the probabilities in two lotteries are the same, and the prize in the first (second) position is the same in the two lotteries, then you must prefer whichever lottery has the better prize in the second (first) position (and if you are indifferent between those, then you must be indifferent between the lotteries).
The third is the *better chances* axiom.

(L3) $(x)(y)(p)(p)[(p > p) \rightarrow (xRy \leftrightarrow L[p,x,y]\ RL[p, x,y])]$

Intuitively: whenever you can choose between two lotteries the same in their prizes, you must prefer the lottery that assigns the higher chance to the better prize.
Finally, there is the reduction of compound lotteries.
Let $d = ab + (1–a)c$.

(L4) $(x)(y)(a)(b)(c)\{L[a,L[b,x,y], L[c,x,y]] \sim L[d,x,y]\}$

Here we deal with complex lotteries, whose prizes are other lotteries. The idea of the axiom is that what matters to the preference ranking of these complex lotteries is only the final chances of getting the various ultimate prizes.
The Expected Utility Theorem follows from these axioms. Rather than proving it, [4] I will just explain what it says. I need three bits of terminology. First, we say that a utility function, u, *represents* a preference ordering R iff:
$(x)(y)[xRy \leftrightarrow u(x)$;eg $u(y)]$
That is, the function assigns greater numbers to preferred prospects, and the same number to any two prospects ranked together.
Second, we say that u is *expectational* iff:
$(x)(y)(p)\{u(\ L[p,x,y]) = pu(x) + (1–p)u(y)\}$
That is, an expectational utility function assigns to each lottery the expected utility of that lottery.

Third, a function *f* is a *positive linear transformation* of a function *g* if there is a real number *b* and a positive real number *a* such that
(x) [$f(x) = a(g(x)) + b$]
end p.159

Now I can state the Expected Utility Theorem. It says that if a preference relation, *R*, satisfies the axioms, then (i) there is some expectational utility function that represents *R*, and (ii) every expectational utility function that represents *R* is a positive linear transformation of every other. The proof of the theorem shows how to construct such a function, given a preference relation that satisfies the axioms.
How to Understand Decision Theory's Advice
Suppose that Liz's preferences conform to the axioms of decision theory. In that case, decision theory constructs an expectational utility function for her. The function will always assign a higher number to the option she prefers, and it will assign to each chancy option the expected utility of that option. This means that Liz never has to *try* to maximize her expected utility. If her preferences conform to the axioms, then the maximization of her utility will take care of itself (as long as she chooses what she prefers!). If her preferences do not conform to the axioms, then the construction of an expectational utility function for her will fail.
It is very misleading, at the least, to say that the advice of decision theory is always to maximize our expected utility in our choices. Insofar as standard decision theory has any normative judgments to make, any advice to give, it is best to think of it as telling us to conform our preferences to its axioms. The axioms are plausibly thought of as constraints of coherence. Decision theory may therefore be thought of as a kind of coherence theory of practical rationality.

2. Social Contract Theory

The contractualist tradition in ethics and political philosophy is sometimes thought to be traceable to the famous Prisoners' Dilemma (PD). There is probably a great deal more to contractualism than can be found in the dynamics of PD, but there is still a lot to the thought. PD is a *game*, represented by a formal model of games, and game theory is typically considered to be a separate (though related) field from decision theory. Games involve strategic interaction, which means that when I am in a game part of my reasoning about what to do involves my reasoning about what you will do, especially including my reasoning about your *reasoning* about what to do. Since I may expect your reasoning about what to do to include some reasoning about how I will be reasoning, game theory includes
end p.160

a type of consideration, potentially very knotty, that does not appear in (what are called) ordinary decision problems.

Prisoners' Dilemma (see also Bicchieri, chap. 10, and Sorensen, chap. 14, this volume) is so famous that I will just include a brief reminder of how it works (see table 9.1). I choose the row, you choose the column; the first payoff listed in each cell is the payoff to me (in utility), and the second is the payoff to you. The game is named after a story in which you and I are prisoners. Each of us can earn a benefit by *Defecting* (choice *D*), and each of us does worse if the other Defects. We both do better if we both *Cooperate* (choice *C*) than we do if we both Defect. However, we cannot influence each other's choices. According to the standard analysis, each of us will Defect, because Defecting *dominates* Cooperating. That is, no matter what you do, I reap a higher utility Defecting than Cooperating, and no matter what I do, you reap a higher utility Defecting than Cooperating. So although both of us prefer the upper left corner of the matrix to the lower right, we will in fact play the moves that put us in the lower right. [5]

Table 9.1 PD

	You	
	C	D
Me C	3,3	0,4
Me D	4,0	1,1

Thomas Hobbes is sometimes said to have understood social relations in the state of nature as a PD game, in which Defecting amounts to making war on one's neighbors and Cooperating amounts to being peaceful. Brian Skyrms (1996) has argued that Hobbes's idea of the state of nature is better modeled as a game of Chicken, which differs from PD in that each player prefers the state in which he Cooperates and the other Defects to the state in which both Defect. Being dominated by one's neighbor in the state of nature is better than fighting in the war of all against all, according to Skyrms's Hobbes (table 9.2). In either case, the general idea is that morality might be thought of as a scheme of rules whose point is to allow us to reap the collective advantage of playing <*C, C*> in a cooperation problem instead of <*D, D*>. [6]

Table 9.2 Chicken

	You	
	C	D
Me C	3,3	1,4
Me D	4,1	0,0

Suppose that you and I could, somehow, together choose rules that both of us would follow in the future. A rule instructing each of us to Cooperate in Prisoners' Dilemmas (or Chicken games) when we face them together would certainly be an attractive rule. If it

were possible to be sure that we both would follow them, you and I would both benefit by agreeing on a cooperation rule.

In some of John Rawls's early work, the choice of principles of justice for a society is thought of along these same lines. [7] Rawls imagines that members of the society come together to choose very general principles that will organize the basic structure of their social lives. Different principles would be advantageous to different people, but as in Chicken and PD, a solution in which people generally cooperate will be superior from everyone's perspective to a solution in which no rules of justice regulate interactions. Rawls suggested that we think of the choice of principles of justice as being made from behind a "veil of ignorance" masking from each party the knowledge of which schemes would be to his or her own advantage. He argued that under such conditions, parties would agree to whichever scheme would maximize the advantage of the least well off. More precisely, Rawls sketched an index of social goods, measuring in a rough way how well off are people or groups in society, and argued that the parties behind the veil of ignorance would choose a principle of justice allowing inequalities with respect to this index only to the extent that a greater equality could be had only by making *all* worse off. John Harsanyi (1976) suggested instead that parties behind a veil of ignorance would reason as utilitarians. Supposing myself to be equally likely to be in any one of the various social positions, I would maximize my expected utility by maximizing the average utility (in the decision theorist's sense) of the members of the society. If the index of social goods is also a utility function for the members, then all parties will prefer the principles that maximize the average index.

Why assume that I am equally likely to occupy any one of the various social positions? Maybe a principle of insufficient reason: since nothing favors my turning out to be a rich industrialist rather than an unemployed farmhand, nor do I have any reason to expect to be in any one of the positions rather than any other, I should just assign them all equal chances. This method doesn't seem *un*reasonable, but neither does it appear to be rationally required.

As a matter of pure decision theory, there doesn't seem to be any very good justification for Rawls's preferred choice of principles. In his later work, and beginning already in *A Theory of Justice*, Rawls abandoned the idea that principles of justice could be grounded in a special application of the theory of rational choice.

end p.162

Gauthier's Rationalist Contractualism

I said earlier that decision theory will typically conflict with morality *if* moral requirements are what common sense presents them as being. David Gauthier's contractualism might be seen as closing the gap between decision theory and commonsense morality. [8] On the one side, Gauthier argues that decision theory must be modified by adding certain constraints. The modified theory still does not match commonsense morality, but it is a lot closer. Then, on the other side, Gauthier argues that where commonsense moral intuitions clash with the dictates of the modified theory of rational choice, it is the intuitions that must give way; they have no practical import.

Suppose that you and I face each other in a game of PD. We will each *Defect*, with the result that we will each reap a utility of one. But now suppose that we could meet in advance and agree upon some strategy, with the agreement determining the outcome. Surely, then, we would agree on mutual *Cooperation*. Agreeing on the pair of strategies <*C, C*> has the following features: it is preferred by both of us to the alternative of reaching no agreement, and there is no alternative that is preferred by both of us. Gauthier's suggestion is that the hypothetical contract, the contract we *would* settle on if we were choosing the outcome together, represents morality.

Two further elements of the theory separate it from other contract theories. First, Gauthier says that the hypothetical contract is rationally binding. In this respect his theory is Kantian. [9] Second, Gauthier provides an ingenious method for finding the right contract given the utility payoffs to parties of the various alternatives.

Why Is the Hypothetical Contract Rationally Binding?

The question becomes more pressing when we think about what is at stake. *We* are the parties to the hypothetical contract. The contractual position is not occupied by imaginary representatives of social groups, or by ideal observers, but by the very people whom it binds in everyday life. If you and I were involved in a PD, we might very well notice that *if* we were to choose together which outcome would prevail, we *would* choose the cooperative outcome. But, we might each think, we are not, in fact, going to get together to choose an outcome. How is the agreement we *would* make at all relevant to our actual choice? Gauthier considers Hobbes's Foole (who sayeth in his heart that there is no justice): [10] "The Foole rejects what would seem to be the ordinary view that, given neither unforeseen circumstances nor misrepresentation of terms, it is rational to comply with an agreement if it is rational to make it" (Gauthier 1986 , 165). The Foole is,

end p.163

in Gauthier's terminology, a *straightforward maximizer;* that is, he will always act so as to maximize his own utility, given the strategies that others adopt. Though he knows that the rational *agreement* in PD is for both to Cooperate, the Foole will Defect, because Defecting brings a higher utility than Cooperating. By contrast, a *constrained maximizer* is defined as "someone who is conditionally disposed to base her actions on a joint strategy or practice should the utility she expects were everyone so to base his action be no less than what she would expect were everyone to employ individual strategies" (167). Gauthier's distinctive claim is that constrained maximization is rational.

The claim is startling from the perspective of decision theory. Forgoing the higher utility cannot be rational (indeed, as I will explain, it is impossible). In defense of the startling claim, Gauthier notes first: "Constrained maximizers canobtain co-operative benefits that are unavailable to straightforward maximizers, however farsighted the latter may be." [11] True enough. It is clear that constrained maximizers get more utility in PD than straightforward ones. But so what? Arguably, I can myself be or become a constrained maximizer just by deciding that it is preferable to be one. On the other hand, it is quite clear that I cannot turn you into a constrained maximizer merely by my preferences. Imagine that whether a person is a constrained or a straightforward maximizer is clear to all those who interact with her. Perhaps the disposition to maximize straightforwardly turns your hair blue, while constrained maximizers have bright red hair. Then constrained

maximizers will cooperate with each other, but they will not cooperate with straightforward maximizers (by definition, above). Straightforward maximizers will not cooperate with anyone. In such an environment, constrained maximizers will do much better (they will regularly reach outcomes they prefer) than straightforward maximizers. [12] If we assume *transparency*, that is, if we assume that the dispositions to choose of every agent are evident to all, then it is advantageous to be or become a constrained maximizer.

We are not transparent. We may, however, be *translucent*. After all, we are able to observe the cooperative behavior of other people, and their character and dispositions manifest themselves to us in the ordinary course of events. We are not utterly opaque. Might constrained maximization be a better strategy under conditions of translucency? The answer depends on a number of factors. It depends on how transparent we are. More precisely, it depends on how likely constrained maximizers are to misidentify other constrained maximizers, and how likely they are to misidentify straightforward maximizers. It depends also on the relative proportion of the two types in the population. If there are few straightforward maximizers, then the risks of constraint are not so great. Third, it depends on how much extra benefit is reaped by cooperation, and how much is lost by exploitation (you are exploited when you Cooperate and your partner Defects) and how much is gained by the

end p.164

exploiter. Gauthier (1986 , 176) gives a formula with these factors as parameters, showing under what conditions it is better to be a constrained maximizer.

To say that it is rational *to be a constrained maximizer* is not obviously the same thing as saying that the choice a constrained maximizer would make is the rational choice. Think of a situation artificially contrived in such a way that irrationality is rewarded. It is surely imaginable that a powerful and knowledgeable eccentric could find and reward irrational people for their irrationality. [13] So even when translucency is sufficient for constrained maximization to be beneficial, it does not quite follow that the constrained maximizing choice is rational. Even if Gauthier must concede this point (he does not concede it), though, the result seems important. If the constrained choice is the moral choice, then the conclusion that we are (under not too implausible conditions of translucency) rationally required to *be* or *become* a chooser of the morally required option is interesting in its own right.

How to Find the Rational Point of Agreement

In PD, there is a very obvious outcome to settle on, if we are getting together to agree on an outcome. Both of us prefer <*C, C*> to <*D, D*>, and no other point is an improvement for both of us over <*D, D*>. So there is no incentive at all, for example, for me to agree to Cooperate while you Defect. I prefer to end negotiations and let us revert to the outcome we will reach in the absence of any agreement.

Other games lack such a salient point of agreement, though, and for them there is no argument as decisive as there is for <*C*, *C*> in PD. Here is an example. Suppose Paul and Lucy can each work separately, or they can work together according to any of three arrangements, represented by these utility payoffs: <8, 2> <1, 10>, <5, 4>, with Paul's payoffs written first, and the option of working separately paying <0, 1>. They want to reach agreement, but which agreement is the right one?
Each player has a most preferred outcome, and each can recognize a "baseline": the outcome that will be chosen if no agreement is reached. Every pair will involve some "concession" by one or both players: the difference between the utility for him or her of that outcome and his or her most preferred outcome. We first calculate these concessions in table 9.3 . The sizes of the concessions are artifacts of the arbitrarily chosen utility scales. But the *relative* concessions are not. A relative concession is measured by the yardstick of how much a player would lose by dropping from his or her most preferred option to the fallback, no-agreement option. For Lucy, this yardstick has a "utility length" of
end p.165

9, since her most preferred option gives her a utility of 10 and she gets 1 if they reach no agreement. For Paul it is 8, since his most preferred option gives him 8 and he gets 0 if there is no agreement. Their relative concessions are the concessions divided by their yardsticks (table 9.4). These relative concessions will not change if we transform Lucy or Paul's utilities by some positive linear transformation.

Table 9.3

Outcome	<8, 2>	<1, 10>	<5, 4>
Concession for Paul	8 − 8 = 0	8 − 1 = 7	8 − 5 = 3
Concession for Lucy	10 − 2 = 8	10 − 10 = 0	10 − 4 = 6

Table 9.4

Outcome	<8, 2>	<1, 10>	<5, 4>
Relative Concession for Paul	0/8 = 0	7/8	3/8
Relative Concession for Lucy	8/9	0/9 = 0	6/9 = 2/3

Gauthier claims that the rational agreement is the one that *minimizes* the *maximum* concession. Find the outcome whose maximum relative concession is smallest. Call this point the *Minimax Relative Concession* point, or MRC. [14] That is the rational point of agreement. In our example, <5, 4> is the MRC, since the maximum relative concession in its column is Lucy's concession of 2 3 , and the other columns have higher maximum relative concessions. So, according to Gauthier, it is the solution that it would be rational for Lucy and Paul to agree on.

This is not obvious. [15] But Gauthier has an argument for it. [16] If some point other than the MRC point is chosen, one of the parties may complain that she is being asked to concede too much. The MRC point has this feature: no alternative would require less relative sacrifice by the person asked to sacrifice most, where sacrifice is measured by relative concession.

The Power of Gauthier's Theory

When the story about rational bindingness is combined with the story about how to find a rational agreement, the result is very powerful. It explains why we should cooperate, why it is *irrational* not to cooperate, and it explains what cooperation involves. If successful, the explanation, although it is not apt to explain our intuitive commonsense morality, will manage to explain certain compelling features of commonsense morality, and, even more important, it will explain why and in what sense we have *reasons* to comply with these moral demands. Gauthier's theory is also important, to my mind, for the criticism, discussion, and commentary that it has inspired. I conclude this section by briefly surveying some of the criticisms.

Is Constrained Maximization Possible?

If we understand utility and preference in the standard way of decision theory, it is very hard to see how it is even possible for someone to avoid being a straightforward utility maximizer. Decision theory constructs a utility function for each agent in such a way that (a) a given option will have a higher utility than a second if and only if the agent prefers it to the second, and (b) the utility for each option will be the same as its expected utility. Imagine that Connie is utterly convinced by Gauthier's argument and decides to become a constrained maximizer. She is no longer a straightforward maximizer. So on some occasion she chooses the option with the lower expected utility. But by definition, the option with the higher utility is the one Connie prefers! So, a decision theorist will say instead that the utility function used to represent Connie's preferences was the wrong one. Since she just chose (for instance) to Cooperate with her partner in what appeared to be PD, her actual utility function must be one that assigns a higher utility to Cooperating than it assigns to Defecting. As long as Connie's preferences satisfy the axioms of decision theory, there is a new function that represents her preferences by an expectational utility function. So she is still a straightforward maximizer.

Now, presumably *something* happened when Connie decided (as she thought) to become a constrained maximizer. Her behavior changed; her disposition to cooperate with others changed. If we cannot describe the change by saying that Connie stopped being a straightforward maximizer, how can we understand it instead? Roughly speaking, Connie acquired a tendency to cooperate with others even when cooperation would not maximize the expectation of her *old* utility function. She now *prefers*, it seems, to cooperate, even when defecting has a higher utility according to her old preferences. [17]

Connie has changed utility functions by changing preferences, as people often do. As a result, the game she is now playing is not PD at all. PD no longer models correctly her interaction with a partner. [18]
end p.167

Restricting the Set of Preferences

What might be thought to be most interesting about Gauthier's argument is that it shows how someone might start from purely selfish preferences and become convinced that she should develop the new dispositions (the dispositions to cooperate with other conditional cooperators). It would be preferable, *from her own selfish point of view*, to have the nonselfish tendencies. I do think that this is an interesting feature of the argument. But it is not completely obvious why it is interesting. After all, most people are not utterly selfish. Is it of merely theoretic interest that a perfectly selfish person could be convinced (under the right circumstances) to change her stripes? Or is there some practical import, too?
Suppose we construct for Connie (whom we will not presume to be utterly selfish) a special utility function, one that represents not her full preference structure but only her *selfish* preferences. Let's call this her *s-utility* function. Now we can think of Gauthier's argument as showing that Connie's s-utility will be higher in the long run if she acts as a constrained s-utility maximizer. [19] Why is this interesting? Well, it may show something about some of Connie's other, nonselfish preferences. It may show that some of her preferences to cooperate actually contribute in the long run to the satisfaction of her selfish preferences. In that case, even in her more selfish moments she can reflectively endorse those nonselfish preferences. [20] It is an important bit of moral psychology to show that certain of our "ethical" preferences are supported and endorsed from the point of view of our more selfish preferences.

The Baseline

Let's look at the question of exactly what is to count as a move in the game. For instance, in the story about Lucy and Paul, the various ways they could work together were counted as possible moves, and so was working separately. But we did not consider the "move" of Lucy's murdering Paul and taking all of his money, or enslaving him at gunpoint and forcing him to work for her. That these were not counted as moves was a critical fact in the subsequent determination of the MRC point of the game, since relative concession is measured by using the yardstick of the difference between favorite outcome and fallback position (no agreement). Why do we count *working separately* as the fallback position, rather than *enslaving the other party*, or perhaps *being enslaved by the other party?*

Gauthier rules out things like enslaving or killing other parties by a version of Locke's famous Proviso. [21] Gauthier's version "prohibits worsening the situation of another person, except to avoid worsening one's own through interaction with
end p.168

that person.This, we claim, expresses the underlying idea of not taking advantage" (Gauthier 1988 , 205). Enslaving another worsens his situation, of course, as does killing him. But we need more precision. Polluting (an example much discussed by Gauthier) also worsens the situation of others. Is polluting ruled out as a possible move in any interactive game? No. Polluting does worsen the situation of another, but not "through interaction," as Gauthier understands it. Don Hubin and Mark Lambeth (1991 , esp. 114–16) give a more precise explanation. The Proviso comes into play whenever I do something that makes you worse off than you would have been had I not existed. It does not, however, debar me from all such actions. Suppose that if I don't do it, then I will be worse off than I would have been had *you* never existed. Then the Proviso no longer applies. Understood this way, the Proviso does express the idea that I must not exploit you. [22]

The Proviso allows me to pollute. Gauthier gives this example: if I live upstream from you, I may reasonably throw my wastes into the stream even though you suffer for it, because had you not been downstream from me I would have polluted anyway. On the other hand, I may not start poisoning the river for the sole purpose of getting you to pay me to stop. That would be taking advantage of you (1988 , 211).

There are some serious questions about whether Gauthier's Proviso captures our intuitive idea of natural rights, but for Gauthier's own purposes, that may not matter much. What does matter is that agreements reached from the Proviso's baseline are rational to keep, while agreements reached by negotiating from "inadmissible" starting points are not. Gauthier's argument is that it is not rational to be disposed to comply with predatory, exploitive bargainers. I will be better off, in the long run and given that others are rational, if I am disposed to refuse to comply with or even bargain with exploitive predators. [23]

Concluding Thoughts on Gauthier

Gauthier's grand project of constructing morality from (self-)interests has produced so much interesting and fruitful commentary that even if it turns out to be an utter failure it will have been worthwhile. [24] The apparatus of decision theory features prominently as a tool in the construction. For the reasons given I doubt that some of Gauthier's main conclusions make sense as long as the concepts of decision theory are interpreted in their standard way. Some adjustments of interpretation can help to preserve the coherence of the theory.

In the next sections I consider more significant departures from the standard interpretation of decision theory. What happens if we try to understand decision theory as being about what is good and bad, rather than about what people prefer?

end p.169

3. The Good

It can be useful to think of decision theory not as a theory of what it is rational to do, but rather as a formal theory. By "formal theory" I mean one left uninterpreted, so that we have just a set of axioms, and we can draw out consequences of those axioms by purely formal means. If we can then find interpretations under which the axioms turn out to be plausible, we will then have a reason to believe the theorems, the consequences of the theory.

It is, as I have said, a mistake to think of decision theory proper as a theory of the good for a person. Be that as it may, decision theory viewed as a formal theory may still be quite relevant to straightforward ethical questions about what is good for a person, and also to questions about what is morally good (and right). We may ask whether the axioms of the theory are plausible if they are interpreted as being about which alternatives are better than which *for* one or another person, or about which alternatives are morally better than which.

In the next section I will present a version of Harsanyi's theorem. In the version I'll present, adapted from Broome 1991 , it purports to be an argument from apparently weak premises to the conclusion that (something very like) utilitarianism is true. I will conclude by explaining which of the premises should seem most questionable to those who find utilitarianism unattractive.

Good for Persons

We want to know whether the relation, "at least as good for a person as," satisfies the constraints given in the axioms of decision theory. That is, if we interpret "*xRy*" as meaning "*x* is at least as good for this person as *y* is," do the axioms come out true? We can start with the ordering axioms.

One might raise questions about completeness. It is at least arguable that becoming an astronaut will not be better for you than becoming a concert violinist, nor will it be worse, nor exactly as good. While "at least as good as" is not obviously complete, I will proceed as if it were. There are degrees of completeness. It may be that a very large field of important prospects can be ordered under "at least as good as," in which case what follows will apply to goodness for persons restricted to that field. Since reflexivity is trivially true of "at least as good as," we can turn to transitivity. It has occasionally been argued that the "better than" relation (for persons) is not transitive. [25] But it must be transitive, for "better than" is a comparative. It means "has more good than." And "more" is logically transitive. If intuitions seem to point to the failure of transitivity of "better than," then either (a) the intuitions are misleading, or (b) the intuitive relation has been misnamed.

end p.170

The lottery axioms are trickier. Are they true, in our interpretation? The reduction of compound lotteries may seem doubtful simply on the grounds that in some circumstances, having to sit through a lottery could itself be a bad thing. Suppose you could either have a simple lottery offering you a 1 2 chance at a new philosophy book and otherwise a tuna sandwich, or else a hugely complex lottery with hundreds of sublotteries nested inside, each as prizes in the lotteries in the levels above, but with the same ultimate prizes (and same chances for them). It's better to have the simple lottery, unless you happen to enjoy the gambling experience and have plenty of free time.

We can finesse this problem. For the purposes of formal decision theory, it is not necessary to think of the lotteries as familiar gambling events. They go on "behind the scenes," as it were; you do not experience them at all as events in their own right. So you won't notice the difference between the compound lottery and its formally equivalent simple lottery.

Continuity presents a more serious issue. Suppose that at the moment, Jessie would enjoy eating a banana. Compare the following prospects:

B She gets the banana.
C She continues as she is now, without the banana.
D She dies immediately.

Plausibly those prospects are properly ranked from better to worse for Jessie, *B* better than *C* better than *D*. Now continuity says that there is some lottery between *B* and *D* that is exactly as good for Jessie as *C* is: call it "*L*[*p, B, D*]." What probability is *p?* Many people think that there is no such *p* (that is, no such $p < 1$; when $p = 1$, the "lottery" is strictly better for Jessie). It is somewhat plausible, at least, that there are some goods that are lexically prior to others, in the sense that it is worse to take any chance of losing the higher-ordered good in exchange for any chance of gaining the lower-ordered.

This may be a real problem. But even if it is real, it may be limited. The "at least as good as" relation might satisfy the axioms over a limited domain, and a separate one might satisfy them again over another, "lexically higher" domain. If there are many different orders of goods, then the application of decision theory to the good for a person will be of relatively little use. But if there are only two, say, then the rich structure of decision theory will still be preserved within the domains. There will be a lower-order expectational utility function representing how good things are for a person, and a higher-order one. Utility space would be two-dimensional. [26] That would be interesting, not disastrous for the model.

I am unaware of any particular difficulties in supposing that "at least as good as" satisfies the *better chances* axiom. There is a well-known objection to the *better prizes* axiom, due to Maurice Allais (see Sorensen, chap. 14, this volume). [27] However, this objection is complicated to present. Since I am going to discuss a closely related problem later, when we consider whether the relation "morally at least as good as" satisfies this same axiom, I will put off consideration of *better prizes* until then.

If the axioms are true under the *personal good* interpretation, then so must be whatever theorems follow. In particular, the Expected Utility Theorem must be. That is, if the axioms are true, then for each person there will be an expectational utility function that represents that person's good. Actually, there will be a whole family of such functions, each a positive linear transformation of the others. The representability of a person's good by an expectational utility function is interesting for a couple of reasons. First, it provides a measure of goodness for persons. We can now ask not only whether one option is better for you than a second, but also how much better. Once a standard unit is fixed, that question will have an answer. It would make sense to say that the change in a person's well-being was twice as great this year as it was last year; that ratio, since it is a ratio of differences, would be independent of the choice of scale (just as the ratio of differences in temperature remains constant no matter which scale we use).
While this result is interesting, more significant is the role it plays in the argument of the next section. Together with some similar assumptions about moral goodness, the assumption that the "at least as good as" relation for persons satisfies the axioms of decision theory turns out to imply a strikingly strong conclusion.

Moral Goodness

Things can be better or worse for persons. More controversially, they can also be simply good or bad. Some philosophers doubt that there is any such thing as a "good state of affairs," [28] but we do seem to have some conception of things turning out better or worse from the moral point of view. In preparation for the main argument, I will ask first whether the relation "at least as morally good as" satisfies the axioms of decision theory, and second whether an option must be (morally) at least as good as another if it is at least as good for everyone. If the answer to both is yes, then a third condition will imply a very strong conclusion, namely a form of utilitarianism.
Let's call the "at least as morally good as" relation "*M*". Now there is a very quick argument to the conclusion that *M* does satisfy the axioms. Imagine someone, call her Angela, who is concerned with nothing other than the moral state of the world. For every pair of alternatives, including lotteries, Angela always prefers the one that is morally better and is indifferent between those that are equally good. Then the ordering imposed by Angela's preference relation is also the ordering of states of the world by *M*. Unless Angela is irrational by decision theoretic standards, her preference relation satisfies the axioms; then so does *M*.
end p.172

There is, however, a particular doubt about the *better prizes* axiom that arises specifically in moral contexts. Suppose that option 1 is definitely better than option 2 in some case, and at least as good in every case. Then the axiom says that option 1 is strictly better than option 2. *Dominance* principles, like *better prizes*, seem very plausible. How could option 2 be as good as option 1 if it could not turn out better and might very well turn out worse? It is hard to see, in the abstract, how something like that might happen. But when the

relation in question is a moral relation, there is a special reason to be wary. This reason has to do with fairness, a specifically moral property. Sometimes an option can be fair exactly because it might turn out one way and might turn out the other way. The fairness resides in the openness of the possibilities, and not locally in one possible outcome or the other. Considerations of fairness jeopardize *better prizes*.

Here is an example. [29] Suppose you have a kitten, which you plan to give away to either Talia or Horace. Talia and Horace both want the kitten very much. Both are deserving, and both would take good care of the kitten. You are sure that giving the kitten to Talia is at least as good as giving it to Horace. But you think that would be unfair to Horace. You decide to flip a fair coin: if the coin lands heads, you will give the kitten to Horace, and if it lands tails, you will give the kitten to Talia. We can call this the *fair chance option*. Let's use *h* for "Horace gets the kitten" and *t* for "Talia gets the kitten." Since giving the kitten to Talia is at least as good as giving it to Horace,

(K1) tRh

Better prizes tells us

(L2) $(x)(y)(z)(p)(xRy \leftrightarrow L[p, z, x]\ R\ L\ [p, z, y])$

so it implies this existential instantiation

(K2) $(z)(tRh \leftrightarrow L[1\ 2\ ,\ z,\ t]\ R\ L[1\ 2\ ,\ z,\ h])$

Now (K1) and (K2) together imply

(K3) $(z)L[1\ 2\ ,\ z,\ t]\ R\ L[1\ 2\ ,\ z,\ h]$

But now instantiate the variable *z* with *t:*

(K4) $L[1\ 2\ ,\ t,\ t]\ R\ L[1\ 2\ ,\ t,\ h]$

The first option, $L[1\ 2\ ,\ t,\ t]$, is the option of flipping the coin and giving the kitten to Talia no matter which side turns up. The second option is the fair chance option. Giving the kitten to Talia is *not* as good as the fair chance option, so (K4) is false. So *better prizes* is false, under the present interpretation.

The kitten story seems to be an example of how dominance can fail. We can make this clearer by supposing that giving the kitten (outright) to Talia is just
end p.173

slightly better than giving it to Horace. Maybe Talia will be able to take somewhat better care of the kitten. It is still plausible that the fair chance option is morally better than giving the kitten outright to Talia. But giving the kitten outright to Talia dominates the fair chance option, since it is strictly better in some case and at least as good in every case. This is a problem for *better prizes*. If it is insurmountable, then moral goodness cannot be represented by an expectational utility function.

Before concluding the moral goodness is not expectational, we should take a closer look at the kitten story. The fair chances option is better than giving the kitten to Talia because it is fair. Fairness, I said, does not reside in the outcome but in the chanciness of the lottery itself. Dominance principles seem plausible when it seems reasonable that all of the goodness in a complex must reside in the simple parts of that complex. The fairness of lotteries, though, does not reside in the outcomes of the lotteries.

Or does it? I am not so sure. Compare two outcomes: first, the fair chances option is chosen and the coin comes up tails, so Talia gets the kitten; second, you decide to give the kitten to Talia, and so you do. At the end of the day, has everything turned out the same in the two schemes? Are the outcomes the same? In one sense, obviously, they are: the outcome of each is that Talia gets the kitten. But in the second scheme Horace has a complaint, and in the first he does not. Isn't that a difference? In the second scheme, Horace has been treated unfairly. In the first scheme, there is no unfairness. That seems to be a difference, and it is no strain to say that it is a difference in the outcomes. Indeed, we might say, if everything *really* turns out *just the same* in the end, then neither case could be better than the other.

The point is that we may have been too quick in identifying outcomes. The outcome of the fair chances option is that Talia gets a kitten *fairly*, while the outcome of giving the kitten to Talia outright is that she gets a kitten *unfairly*. A proper representation of the two options ought to represent that difference. So maybe they should be represented like this:

$L[1\ 2\ ,\ t^u\ ,\ t^u\ ]$ and $L[1\ 2\ ,\ t^f\ ,\ h]$

But then the example does not even appear to violate *better prizes*.

The issue is crucial to Harsanyi's argument. It shows up again in a slightly different guise in the question of the soundness of an extra premise we need for that argument. To recap, we are going to use the assumption that *at least as good as* for persons satisfies the decision theory axioms, and the assumption that *M* satisfies the axioms, and then two more premises. The first of these we can call Personal Good: [30]
end p.174

(PG) If *x* is at least as good for each person as *y*, and better for some person, then *x* is morally better than *y;* and if *x* is exactly as good for each person as *y*, then *x* is exactly as good as *y*.

We need one more premise. Following standard practice I will call it Anonymity:

(A) Nobody's good matters more than anyone else's.

Surveying the distribution of good to persons, according to (A), we ought to be able to see how (morally) good the situation is without knowing the identities of the bearers of good. The persons can be just as well left anonymous.

4. Harsanyi's Theorem

John Harsanyi put his theorem in terms of the more standard interpretation of decision theory. He imagined that the preferences of individual members of a society met the constraints of decision theory, and that a "social preference" ordering did too. Adding the analog of (PG) and (A) to connect the members' preferences with the social ordering, he showed that the social preference ordering must be representable by a utility function that is the sum of the utility functions of the individual members. But the conclusion of the argument is more obviously relevant to ethics if it is about what is morally good and what is good for individual people. So as we will interpret the theorem, it says that there is a utility function that represents moral goodness and is the sum of utility functions representing the good of persons. [31]

Step 1: Choice of Scales

Since we have leeway in our choice of utility functions (we may choose from the family related by positive linear transformation), we will choose some that are convenient for the presentation. So first, we will represent the worst prospect for any given person by the utility number 0, and the best by the utility number 1. [32] Let U_1, U_2, etc. be the utility functions representing the good for persons 1, 2, etc. And the utility function representing overall moral goodness will be W. W must assign the same number to any pair of alternatives that are assigned the same number by all of the U_i. [33] Also, (PG) tells us that the alternative in which every
end p.175

person is worst off must be the worst possible alternative, since every distinct alternative will be better for someone and worse for no one. We'll select our moral goodness scale so that this morally worst alternative is assigned 0. This means that
$W(0, 0, 0,, 0) = 0$
where the arguments are, as I said, the utility numbers measuring the good for each person. We complete the choice of a scale for W by stipulating that
$W(1, 0, 0,, 0) = 1$

By (A),
$W(1, 0, 0,, 0) = W(0, 1, 0,, 0) == W(0, 0, 0,, 1) = 1$
That is, all the alternatives that are good to degree 1 for one person and worst for each other are exactly as good as one another (that's Anonymity), and we've stipulated that the degree of moral goodness of these alternatives is one. We'll call these "unit vectors." [34]
Step 2: Multiplication Lemma
What I'll call the multiplication lemma says that for any real number k between 0 and 1,

(ML) $kW(u_1, u_2,, u_n) = W(ku_1, ku_2,, ku_n)$

where the u_i are the utility numbers representing how good the alternatives are for persons 1 through n. That is to say, halving the goodness for each person will half the overall goodness, and generally multiplying the goodness for each person by k will multiply the overall goodness by k. This is hardly obvious. But it is not hard to demonstrate.
Consider the lottery
$L^* = L[k, (u_1, u_2,, u_n), (0, 0,, 0)]$
Because all of the utility functions are expectational, we can work out quickly some definite facts about the utilities of L^*. For example, the goodness of L^* for person 1 must be ku_1, since that is the expected goodness for person 1.
$U_1(L^*) = ku_1$; $U_2(L^*) = ku_2$;; $U_n(L^*) = ku_n$

The overall goodness of the worst case scenario is also 0, so the expected moral goodness of L^* is also a simple matter:
$W(L^*) = kW(u_1, u_2,, u_n)$
But now we may just replace the "L^*" on the left side with the list of utilities representing the goodness of L^* for each person:
$W(ku_1, ku_2,, ku_n) = kW(u_1, u_2,, u_n)$
which is the multiplication lemma.
Step 3: Harsanyi's Theorem
We will show that $W(u_1, u_2,, u_n) = u_1 + u_2 ++ u_n$, that is, that the moral goodness is the sum of the goodness for persons. The proof will be for the case of two persons only; the generalization to many persons is direct but because of our notation it would be very cumbersome.
If we are given u_1 and u_2, we can then consider the lottery $L[1/2, (u_1, 0), (0, u_2)]$ giving equal chances to $(u_1, 0)$ and $(0, u_2)$. Call this lottery $L^=$. Its moral goodness must be equal to its expected moral goodness, so
$W(L^=) = W(u_1, 0)/2 + W(0, u_2)/2$
Its expected goodness for person 1 is $u_{1/}2$, and for person 2 $u_{2/}2$, so
$U_1(L^=) = u_{1/}2$; $U_2(L^=) = u_{2/}2$
The moral goodness of $L^=$ is just a matter of how good it is for each person, so
$W(L^=) = W(u_{1/}2, u_{2/}2)$,
so by the multiplication lemma,
$W(L^=) = W(u_1, u_2)/2$

Now combine the first and last lines to get
$W(u_1, 0)/2 + W(0, u_2)/2 = W(u_1, u_2)/2$
end p.177

so as a matter of algebra,
$W(u_1, 0) + W(0, u_2) = W(u_1, u_2)$
But the multiplication lemma tells us that
$W(u_1, 0) + W(0, u_2) = u_1 + u_2$
So, $W(u_1, u_2) = u_1 + u_2$, which is what we wanted to show.
Remarks on Harsanyi's Theorem
The conclusion of the argument rests on a handful of assumptions. Several of these assumptions are questionable. However, the argument is still remarkable, I think. The assumptions *appear* to be a lot weaker than the conclusion (though of course, collectively they couldn't be weaker than the conclusion!). I want to point out two things. First, we have to revisit Anonymity. Second, I want to elaborate on a crucial step in the argument, the third step, explaining how one of the controversial assumptions figures in that step.
First, for Anonymity to do the work it has to do in the proof, we must in effect assume that we have a clear way to compare how well off I am with how well off you are. If we have such a thing, then Anonymity really does mean that swapping numbers from one position to another on a vector cannot make it better or worse. But if the goodness scales for you and me are really just arbitrary, so that your 0 and my 0 are in no sense "the same degree of welfare," then it is hard to see what justification we could have for supposing that permuted vectors must be exactly as (morally) good as one another. The assumption that we can compare my welfare with yours is a strong one. It does seem reasonably intuitive that we can do it, but it is hard to see how to do it rigorously.
Now to (what I think is) the crucial step, the third one. There we considered a fair lottery between $(u_1, 0)$ and $(0, u_2)$ and said that this lottery must have goodness of $u_1/2$ for person 1 and $u_2/2$ for person 2. We might plug in numbers to make things clearer; let's have $u_1 = u_2 = 2$, to fix ideas. Then these lotteries are good for person 1 (me, say) to degree 1 and for you (person 2) also to degree 1. That means the lottery is exactly as good for you as (1, 1) and also exactly as good for me as (1, 1). Call the (1, 1) outcome the egalitarian outcome, and (2, 0) and (0, 2) the inegalitarian outcomes. Since the lottery between the inegalitarian outcomes is exactly as good for each of us as the egalitarian outcome is, (PG) implies that the lottery is exactly as good morally as the egali
end p.178

tarian outcome. But one might think the egalitarian outcome is better. After all, it is egalitarian! And that is something that can't be said for the outcomes of the lottery. Still, it might be thought that an egalitarian outcome is not better than a fair lottery between inegalitarian outcomes. That person 1 does better than person 2 may not be an objection, as long as person 2 had an equal chance of doing better. The fairness may reside in the whole of the lottery, as it did in the kitten story.
But notice now that the outcome (0, 2) must be exactly as good as the outcome (2, 0). But then *better prizes* tells us that (0, 2) is exactly as good as $L[1/2, (2, 0), (0, 2)]$. This seems

wrong, and for exactly the reason that giving the kitten to Talia seemed intuitively worse than holding a fair lottery between Talia and Horace. And again, *better prizes* is the culprit.

Now as before, we might say that the outcomes are not properly represented. If person 2 ends up on the short side of things *as the result of a fair lottery*, that is not the same outcome as her having had no chance of the better outcome. But here the outcomes are represented only by a sequence of utility numbers. If these numbers do represent goodness for the persons involved, and if fairness counts extra toward the overall goodness of the outcome, then that extra goodness couldn't be goodness for some person. That means that (PG) is false. Maybe fairness is a good that is not a good *for anybody*.

The argument of Harsanyi's theorem makes some strong assumptions about where goodness of an outcome can be located. It assumes that all goodness must be goodness for someone (that's what (PG) says), and also that the goodness of a prospect must reside entirely in the goodness of the prospective outcomes (that's what *better prizes* says). These are dominance principles, and also "antiholism" principles. Goodness of a complex must reside in the parts, and not in their relations to one another. These assumptions may be wrong.

Even if they are wrong, I think Harsanyi's theorem provides an illuminating argument. It gives us one way of thinking about which assumptions are fundamental to utilitarianism. And it shows how these assumptions figure in the overall theory. Considerations of *better prizes* and (PG) may also show something about utilitarianism and may help to explain certain counterintuitive features of the view. Utilitarianism does not recognize the good of equality; it does not recognize how distributive considerations may be important. Its antiholism is traceable to the dominance principles at work in the Harsanyi argument.

end p.179

NOTES

1. I'll follow the terminology of Resnik 1987 , which follows von Neumann and Morgenstern's formulation.
2. See Dreier 1996 for a defense of this interpretation.
3. See Resnik 1987 , chap 4. Resnik's book provides compact and clear explanations of both formal and philosophical topics within and arising out of decision theory.
4. It is proved in Resnik 1987 , 88–98.
5. In this paragraph, and hereafter, I use the capitalized "Cooperate" and "Defect" as names for moves in the game of Prisoners' Dilemma.
6. The structure of the two games depends only on the *ordering* of the outcomes for each player. I have selected the payoffs in such a way that the total utility is greater for the two players combined in the perfectly cooperative outcome, and least in the perfectly uncooperative outcome. This feature will help me make a couple of points later.
7. Beginning with Rawls 1951 , and fading off rather than ending at a sharp boundary line, but in any case most famously in Rawls 1971 .
8. In a number of papers, but culminating in Gauthier 1986 .

9. The theory is Kantian in another respect. Morality is binding on us not merely in virtue of the preferences we happen to have, not merely because or insofar as we happen to care about others, but also in virtue of our rationality.
10. In Hobbes 1968 , chap. 15.
11. Gauthier 1986 , 171. There is another clause, which I omit for brevity.
12. Axelrod 1984 gives a very satisfying and compelling presentation of some computer simulations that demonstrate the superiority, in many different environments, of reciprocal cooperation.
13. Some people think that Newcomb's Predictor acts in exactly this way. See Lewis 1985 .
14. See Gauthier 1986 , chap. 5, especially 136–45.
15. In fact, it bucks the tradition of Nash, Zeuthen, and Harsanyi, who argue for a different method of determining the rational bargain. See Luce and Raiffa 1989 , 124ff.
16. This argument is not easy to find. Here I am working from Gauthier 1986 , 137–39.
17. In this paragraph I have not capitalized "cooperate," "defect," or their cognates. That's because the capitalized terms are technical terms; they are the names of moves in the theoretic game of PD, as I explained in n. 5.
18. See Blackburn 1998 , chap. 6, for a similar diagnosis of what has happened when somebody becomes a constrained maximizer.
19. See Gauthier's many discussions of "mutual unconcern" and "non-tuism." Gauthier does not construe his own argument as I am suggesting here. For an interesting discussion of Gauthier's own use of the assumption of "mutual unconcern," see Thomas 1988 .
20. Her attitude toward her own cooperative attitudes might be something like the attitude that Hume ([1739] 1978) thought we should have toward our "artificial virtues," in book 3, part 2 of the *Treatise*.
21. See Locke's *Second Treatise of Government* ([1690a] 1952), chap. 5, paras. 27 and end p.180

33: "Whatsoever [a man] removes out of the state that nature has provided, he has mixed his labor with, and joined to it something that is his own, and thereby makes it his property.For this labor being the unquestionable property of the laborer, no man but he can have a right to what that is once joined to, *at least where there is enough and as good left in common for others*." The Lockean Proviso, so-called by Robert Nozick (1974 , 175ff.), is the portion italicized.
22. Or at least it approximates that idea. See Hubin and Lambeth 1991 for a discussion of complications.
23. This claim may seem dubious. It is discussed by several critics in part 2 of Vallentyne 1991 .
24. Vallentyne 1991 is a collection of uniformly high quality and remarkably broad interest.
25. See, e.g., Temkin 1996 .
26. As explained in Hausner 1954 .
27. In Allais 1953 .
28. See, e.g., Thomson 1997 and Foot 1988 .

29. The example is modeled after one given in Diamond 1967 .
30. Following Broome 1991 , especially 165. Broome's book provides a superb philosophical treatment of the application of decision theory to the ethical concept of goodness, in far more depth than I manage in this chapter.
31. Aside from this difference in interpretation and a few minor differences, the proof here follows Resnik 1987 , 197–200.
32. This assumes that there are best and worst possibilities for persons, which may not be true. The proof still works without the assumption, but it is more cumbersome.
33. For this reason, I am going to use the functor *W* indiscriminately as a functor of one place, namely the option to be morally evaluated, and a functor of many places, namely the utilities representing how good the option is for each person. I hope this reduces confusion rather than aggravating it.
34. It is sometimes asked whether we can be sure all the unit vectors exist. The question is a little more worrisome in the context of Harsanyi's original argument, where the locations on the vectors represent people's *preferences*. In our context, there could be a real worry if we insisted that each alternative be actual, but there is no particular reason so to insist.

Chapter 10
RATIONALITY AND GAME THEORY

Cristina. Bicchieri

Game theory (see also Danielson, chap. 22, this volume) aims to understand situations in which decision makers interact. Chess is an example, as are firms competing for business, politicians competing for votes, jury members deciding on a verdict, animals fighting over prey, bidders competing in auctions, threats and punishments in long-term relationships, and so on. What all these situations have in common is that the outcome of the interaction depends on what the parties jointly do. Decision makers may be people, organizations, animals, robots, or even genes. The theory of rational choice is a basic component of game-theoretic models. This theory has been mostly criticized from a descriptive viewpoint, arguing that it requires far too many calculating capabilities from ordinary beings that use at most simple heuristics. Few, however, have given a close and critical look at how a normative theory of rational choice fares in interactive decision contexts. What does it mean to be rational when the outcome of one's action depends upon the actions of other people and everyone is trying to guess what the others will do? In social interaction, rationality has to be enriched with further assumptions about individuals' mutual knowledge and beliefs, but these assumptions are not without consequence. In what follows I shall spell out these extra assumptions, and see whether they are sufficient to lead the players to coordinate upon outcomes that are mutually acceptable. Since the issue of whether rational choice theory is an adequate foundation for game theory is raised in the
end p.182

context of noncooperative games, I will restrict my attention to this class of games.

Rational Choice

The theory of rational choice's central assumption is that a decision maker chooses the best action available according to her *preferences*. The content of preferences is unrestricted. Agents' preferences may be selfish or altruistic, self-defeating or even masochistic. Preferences mirror values and dispositions that are beyond the pale of rationality. What is required is that preferences are well behaved, in the sense of fulfilling certain formal conditions I list below. If preferences are well behaved, they can be represented by utility functions, and rationality consists in maximizing one's utility, or finding the maximum value of one's utility function.

When we model a choice situation, we must be careful to follow a series of steps. First, we must define a set of feasible actions A. These are the actions an agent knows are available to him. If an action is available, but one does not know about it, then we do not include it in the set of feasible actions. In the simplest case, each action has a unique, certain consequence. We must then list the set C of the consequences of actions. Since an agent is assumed to have preferences over consequences, the third step involves introducing a binary weak preference relation R on the set C that is complete (for any x, yC either xRy or yRx), antisymmetric (if xRy and yRx, then $x \sim y$), reflexive (for all xC, xRx) and transitive (if x, y, z C, xRy and yRz, then xRz). ($\sim$ denotes indifference, P (used below) denotes strict preference.) The preceding assumptions about the weak preference relation on C identify a qualitative preference structure. A *representation* provides a correspondence between this structure and properties of real valued functions based on C. In other words, we want a numerical representation for the preference relation, such that higher numbers correspond to more preferred consequences. Utility functions are just such representations. An agent's preferences will then be represented by a utility function U: $C \rightarrow \mathbf{R}$, such that xRy iff $U(x)$;eg $U(y)$, for any x, yC. Since in this case the utility function only conveys ordinal information, any increasing function of U can represent an agent's preferences. The final step consists in defining rational choice proper. If each action has a unique consequence, we may introduce a consequence function f: $A \rightarrow C$ that associates a consequence with each action. A *rational agent* will choose an action a^* that is feasible and optimal, in the sense that, for all aA, $f(a^*)$ R
end p.183

$f(a)$. Alternatively, we may say that a rational agent will act *as if* maximizing the utility function $U(f(a))$.

A far more common case is one in which an action may have one of several possible consequences. A situation of *risk* is one in which the probabilities with which the consequences occur are objective and known to the decision maker. Then the consequence function f is stochastic and is known to the agent (i.e., for each action a, $f(a)$ is a probability distribution on C). Suppose action a has two possible consequences, x and x, which occur with probability p and $(1–p)$, respectively. Choosing action a is like choosing a lottery that gives prize x with probability p, and prize x with probability $(1–p)$. We assume agents to have preferences over such lotteries. If preferences are complete, transitive, and satisfy a number of other conditions (von Neumann and Morgenstern 1944), they can be represented by the expectation of a real-valued utility function U: $C \rightarrow \mathbf{R}$

(unique up to a positive linear transformation) such that, for any two lotteries a and b, aPb iff $\Sigma_{xa}\, p(x)U(x) > \Sigma_{yb}\, p(y)U(y)$. A rational agent will choose an action (lottery) a^* that maximizes the expected value of a von Neumann–Morgenstern utility function.
If the stochastic connection between actions and consequences is not given, the decision maker will have to make some hypotheses about which states of nature might obtain, since the consequence of an action will depend upon which of a number of mutually incompatible states of nature is realized. The ingredients of choice under uncertainty are thus consequences and states of nature. A state of nature indicates, for each feasible act, what the consequence will be. The consequence function f is now defined as f: $A\ X\ S \rightarrow C$, so when we write $f(a, s)$, we mean the consequence of taking action a if state s is realized. The desirability of the consequence depends neither on the act, nor on the state of nature that makes it possible. An agent is assumed to consider a state space S, and to act in a manner consistent with having a subjective probability measure $\boldsymbol{p}$ over S (Savage 1954). $\boldsymbol{p}$ is an additive subjective probability measure over S. A subjective expected utility representation for a system of preferences exists if preferences satisfy some necessary conditions formulated by Savage (1954). In this case, there exists a real-valued utility function U: $C \rightarrow \boldsymbol{R}$ such that, for any two actions a and b, aPb iff $\Sigma_{sS}\, \boldsymbol{p}\,(s)U(f(a, s)) > \Sigma_{sS}\, \boldsymbol{p}\,(s)U(f(b, s))$. A rational agent will select an action a^* that maximizes the expected value of $U(f\,(a^*, s))$, relative to the probability measure.
In all three cases, rationality is identified with (expected) utility maximization, and this presupposes that individuals have (or behave *as if* they have) utility functions, that is, that they have well-behaved preference orderings over alternatives. In game-theoretic models, von Neumann–Morgenstern utility functions are the most commonly used. Such utility functions are unique only up to a positive linear transformation, which means that for any set of outcomes there is an in
end p.184

finite number of numerical functions, mutually related by positive linear transformations, each of which represents an agent's utility as well as any other. A consequence of using von Neumann–Morgenstern utility functions is that it becomes impossible to make interpersonal utility comparisons. To understand why, imagine trying to compare two temperatures that are expressed in Celsius and Fahrenheit measures, without having a scale. In noncooperative games this is not a problem, since interpersonal utility comparisons are never brought into play. In evolutionary games instead, interpersonal utility comparisons may be needed. This is especially true in models of cultural evolution, the process through which behavioral patterns that are not genetically inherited are transmitted from one generation to another, or spread from one group to other groups. One such transmission process is imitation, which might involve comparing one's payoffs with another's, and switching strategies if the other's payoffs are better than one's own. Interpersonal comparisons may be a completely subjective matter, but if this were the case, we could not talk of a selection process that favors better, more efficient strategies over less efficient ones. For selection to work, we need at least intersubjective agreement on which strategies are “most successful,” and this involves interpersonal comparisons among payoffs. If payoffs are expressed in von Neumann–Morgenstern's utilities, however, such comparisons are impossible.

Strategic Interaction

In a strategic interaction, the outcome of an action depends, among other things, upon the actions of other agents. Other agents have plans, preferences, and beliefs, and unless one is certain which action will be chosen by another agent, one will have to form beliefs about other agents' possible choices, and even beliefs about the expectations that may guide another agent in choosing a particular action. Whereas rational choice is relatively straightforward in individual decision making, it becomes more complicated in a strategic decision context. When we abstract from the particulars of such contexts and formally describe the strategic interaction, we have a game.

Any strategic interaction involves two or more decision makers (players), each with two or more ways of acting (strategies), such that the outcome depends on the strategy choices of all the players. Each player has well defined preferences among all the possible outcomes, enabling corresponding von Neumann–Morgenstern utilities (payoffs) to be assigned. A game makes explicit the rules gov

end p.185

erning players' interaction, the players' feasible strategies, and their preferences over outcomes.

One way of representing games is in *normal form*. A normal form game is completely defined by three elements: a list of players $i = 1,, n;$ for each player i, a finite set of pure strategies S_i ; a payoff function u_i that gives player i's payoff $u_i(s)$ for each n-tuple of strategies $(s_1,, s_n)$, where u_i : X>[md4]n[mu4][mu6]>i=1[md6] Si → $\boldsymbol{R}$. All players other than some given player i are customarily denoted as $-i$. A player may choose to play a pure strategy, or instead he may choose to randomize over his pure strategies; a probability distribution over pure strategies is called a *mixed strategy* and is denoted by σ_i . The pure strategies over which a player randomizes are called the *support* of the resulting mixed strategy. Each player's randomization is assumed to be statistically independent of that of his opponents, and the payoffs to a mixed strategy are the expected values of the corresponding pure strategy payoffs.

The 2×2 matrix in figure 10.1 depicts a two-player normal form game: each player picks a strategy independently, and the outcome, represented in terms of players' payoffs, is the joint product of these two strategies. Often the normal form representation is interpreted as a situation in which the players choose simultaneously, without any knowledge of each other's choice. This temporal interpretation, though intuitive, is not quite correct. Since the normal form is not meant to convey any information about players' knowledge of each other's moves, or their temporal order, it is best to interpret it as a list of outcomes, one for each possible combination of strategies the players might choose. If we want to model the information (or lack thereof) that players have about each other and the order in which they move, we have to choose another type of representation. I will return to this point later.

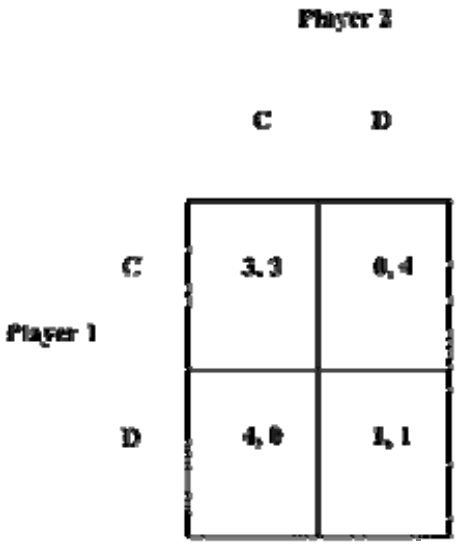

Figure 10.1

The game of figure 10.1 is one of *complete information*, in that players are assumed to know the rules of the game (which include players' strategies) and other players' payoffs. If players are allowed to enter into binding agreements before the game is played, we say that the game is *cooperative*. *Noncooperative games* instead make no allowance for the existence of an enforcement mechanism that would make the terms of the agreement binding on the players. If preplay negotiation leads to an agreement, and there is no enforcement mechanism available, we should ask whether the agreement is self-enforcing, by virtue of being in the best interest of each player to adhere to it.

Nash Equilibrium

Expected utility maximization has always been a building block of game theory, but for many decades game theorists have paid little attention to the link between rational choice and strategic interaction, or how the outcome of strategic interaction can be derived from rational choices. In a well-known passage of their book, *Theory of Games and Economic Behavior*, von Neumann and Morgenstern say that rational players who know (i) all there is to know about the structure of the game they are playing, (ii) all there is to know about the beliefs and motives of the other players, (iii) that every player is rational, (iv) that every player knows (i)–(iii), (v) that every player knows (i)–(iv), and so on, will be able to infer the optimal strategy for every player. In that case, each player will behave rationally by maximizing his expected utility conditional on what he expects the others to do.

This idea that rational players will always jointly maximize their expected utilities, or play a *Nash equilibrium*, could rightly be called the “central dogma” of game theory. Nash equilibrium (Nash 1951) is the standard solution concept for noncooperative games. Informally, a Nash equilibrium specifies players' actions and beliefs such that (i)

each player's action is optimal given his beliefs about other players' choices, and (ii) players' beliefs are correct. Thus an outcome that is not a Nash equilibrium requires either that a player chooses a suboptimal strategy, or that some players "misperceive" the situation.

More formally, a Nash equilibrium is a vector of strategies ($\sigma^*_1, \ldots, \sigma^*_n$), one for each of the n players in the game, such that each σ^*_i is optimal given (or is a *best reply* to) σ^*_{-i}. Note that optimality is conditional only on a fixed σ_{-i}, not on all possible σ_{-i}. A strategy that is a best reply to a given combination of the opponents' strategies may fare poorly vis-à-vis another strategy combination. Consider for example figure 10.2 .

end p.187

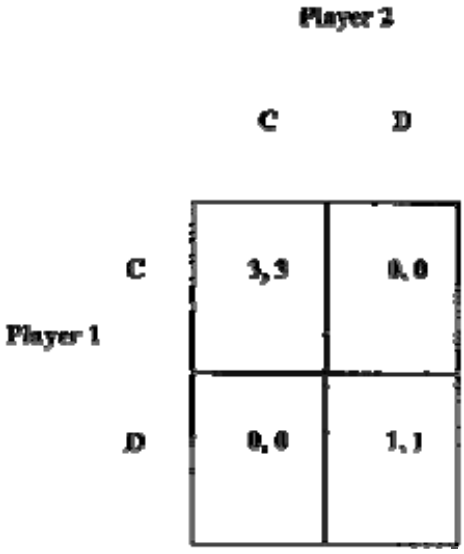

Figure 10.2

In this game, there are two Nash equilibria, (*C, C*) and (*D, D*), but a player's strategy *D*, for example, is optimal only if the other player is expected to play *D*, too. If the other player were to play *C* instead, the choice of *D* would lead to a poor outcome. A common interpretation of Nash equilibrium is that of a self-enforcing agreement. Were players to agree in preplay negotiation to play a particular strategy combination, they would have an incentive to stick to the agreement only in case the agreed upon combination is a Nash equilibrium. In the case of a *strict* Nash equilibrium, any deviation from the equilibrium strategy nets a player an inferior payoff. If the equilibrium is not strict, however, a deviation from equilibrium play may earn a player the same payoff as the equilibrium strategy. In the latter case, the incentive to follow the Nash equilibrium is less strong. The lack of a strong incentive to play one's part in a Nash equilibrium is particularly obvious in the case of mixed strategy equilibria, which, as we shall see, are never strict. Consider the game in figure 10.3 .

This game has no Nash equilibrium in pure strategies, but Nash proved that—provided certain restrictions are imposed on strategy sets and payoff functions—a game has at least an equilibrium in mixed strategies. Nash's result generalizes von Neumann's theorem

(1928) that every noncooperative game with finitely many strategies has an equilibrium in mixed strategies.
Suppose 1 plays (4/9 *a*, 5/9 *b*). Then if 2 chooses *c*, her expected utility is 4(4/9) + 7(5/9)= 17/3. If 2 chooses *d*, she nets 9(4/9) + 3(5/9) = 17/3. So if 1 randomizes between *a* and *b* with probabilities (4/9, 5/9), 2 is indifferent between *c*, *d*, or a lottery in which she chooses *c* with probability *p* and *d* with probability (1–*p*). Suppose 2 chooses (4/7 *c*, 3/7 *d*). In this case 1 nets 48/7 if he plays *a*, and 48/7 if he plays *b*. Hence 1 is indifferent between *a*, *b*, and any lottery (*ap*, *b*(1–*p*)). The combination (4/9 *a*, 5/9 *b*), (4/7 *c*, 3/7 *d*) is a mixed strategy Nash equilibrium.
end p.188

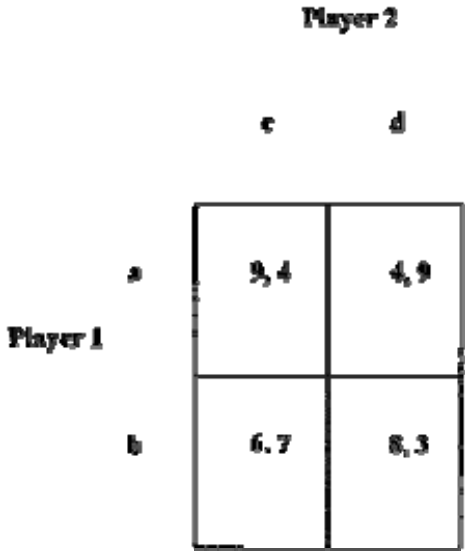

Figure 10.3

In a mixed strategy equilibrium, the equilibrium strategy of each player makes the other indifferent between the strategies on which he is randomizing. For example, if 1 were to know that 2 randomizes with probabilities (4/7, 3/7), any of his strategies (pure or mixed) would be a best reply to 2's choice, and conversely, were 2 to know that 1 randomizes with probabilities (4/9, 5/9), any of her strategies, pure or mixed, would be a best reply. Paradoxically, if players agree to play a mixed strategy equilibrium, they have no incentive to play their part in the equilibrium. A mixed strategy equilibrium is a self-enforcing agreement only in the weak sense that—given the other players' equilibrium behavior—each player is indifferent between all the strategies (and lotteries over these strategies) in the support of her equilibrium mixed strategy.
There are, however, more serious questions raised by the Nash equilibrium concept. Ken Binmore (1987 , 1988) has argued that there are two possible interpretations of Nash equilibrium. According to the *evolutive* interpretation, a Nash equilibrium is an observed regularity. Players know the equilibrium, and test the rationality of their behavior given this knowledge acquired from experience. The players (and the game theorist) can accordingly predict that a given equilibrium will be played, since they are accustomed to

coordinate upon that equilibrium and expect (correctly) others to do the same. According to the more commonly adopted *eductive* interpretation instead, a game is a unique event. In this case it makes sense to ask whether players can deduce what others will do from the information available to them. The players (and the game theorist) can predict that an equilibrium will be played just in case they have enough information to infer players' choices. The standard assumptions game theorists make about players' rationality and knowledge should in principle be sufficient to guarantee that an equilibrium will obtain. The following assumptions are standard:
end p.189

CK1. The structure of the game, including players' strategy sets and payoff functions, is common knowledge among players.
CK2. The players are rational (i.e., they are expected utility maximizers) and this is common knowledge.

The concept of *common knowledge* was introduced by Lewis (1969), and later formalized by Aumann (1976). Simply stated, common knowledge of *p* among a group *G* means that each member of *G* knows *p*, and each knows that each knows *p*, and so on ad infinitum. Common knowledge of rationality, preferences and strategies may facilitate the task of predicting an opponent's strategy but, as I argued elsewhere (Bicchieri 1993), it does not guarantee that the resulting prediction will be correct.
Consider a game that has a unique Nash equilibrium in pure strategies (figure 10.4). Can the players infer what other players will do from CK1 and CK2? Here player 1 has two pure strategies, *A* and *B*, and player 2 has three pure strategies, *a*, *b*, and *c*. There is a unique Nash equilibrium in pure strategies, (*B*, *a*), but it is not evident that players can infer that it will be played by reasoning from CK1 and CK2. As an example of how players may reach a conclusion on how to play, consider the following argument by player 1. "If player 2 believes that I will play *A*, then it is optimal for her to pick *c*. And why would she think I play *A*? Well, she must believe that I expect her to play *b*, to which *A* is a best reply. And why would I expect her to play *b*? I would (she will think), if I were to believe she expects me to play *B*" It is easy to verify that such a chain of reasoning can justify the choice of *any* strategy for both players.
The concept of Nash equilibrium embodies a notion of individual rationality, since each player's equilibrium strategy is a best reply to the opponents' strategies, but unfortunately it does not specify how players come to form the beliefs about each other's strategies that support equilibrium play. Beliefs, that is, can be internally consistent but fail to achieve the interpersonal consistency that guarantees that an equilibrium will be attained. Bernheim (1984) and Pearce (1984) have argued that assuming players' rationality (and common knowledge thereof) can guarantee only that a strategy will be *rationalizable*, in the sense of being supported by internally consistent beliefs about other players' choices and beliefs. But a combination of rationalizable strategies may not constitute a Nash equilibrium. In the game depicted in figure 10.4, all six combinations of strategies are rationalizable, yet only one of them is an equilibrium. The fact that a Nash equilibrium is

always a combination of rationalizable strategies is of no help in predicting it will be played.
Note that there are strategies that are not rationalizable, in the sense of not being supported by coherent beliefs. Consider the game in figure 10.1. Here the *C*-strategy is never rationalizable, since each player is always better off by choosing
end p.190

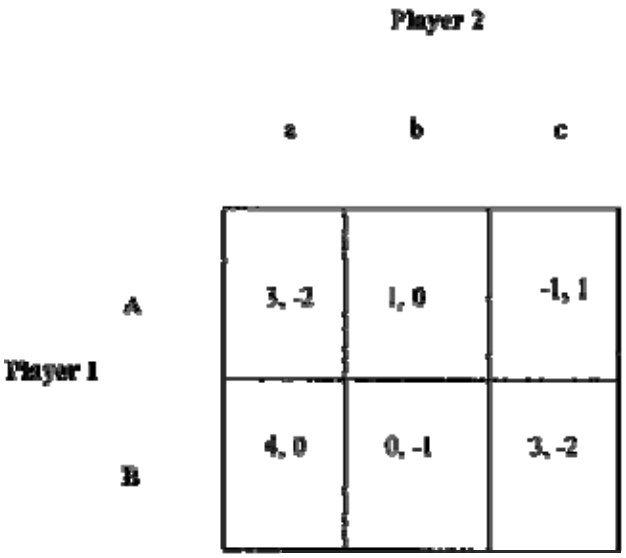

Figure 10.4

strategy *D*, whatever the other does. And each player is able to see that the other player, if rational, will never choose strategy *C*. In this case, CK1 and CK2 will lead the players to predict the outcome of the game.
In a Nash equilibrium, the optimality of a strategy is conditional only on a fixed σ_{-i}, not on all possible combinations σ_{-i}. In the game of figure 10.1, however, the Nash equilibrium strategies (*D*, *D*) are also optimal with respect to any strategy choice of the opponent. Whatever player 1 does, player 2 is better off by choosing *D*, and the same is true of player 1. We say that a strategy s_i is *strictly dominated* by another strategy t_i if, for every choice of strategies of the other players, *i*'s payoff from choosing t_i is strictly greater than his payoff from choosing s_i. In our example, *C* is strictly dominated by *D* for both players. A strictly dominated strategy is never rationalizable, since the belief that a player plays it is inconsistent with common knowledge of rationality. We say that s_i is *weakly dominated* by t_i if, for every choice of strategies of the other players, *i*'s payoff from choosing t_i is at least as great as *i*'s payoff from choosing s_i. Note that weakly dominated strategies are rationalizable, since there always exists an opponents' strategy combination to which a player's weakly dominated strategy is a best reply.
When a game has a unique Nash equilibrium, we can predict that it will be played if we are able to show that players, armed with common knowledge of rationality and of the structure of the game, will infer the Nash solution. If players have dominated strategies, CK2 entails that they will eliminate them, and this is common knowledge (we assume

that the consequences of CK1 and CK2 are common knowledge, too). Often after we have eliminated strictly dominated strategies for one player, we may find that there are now strictly dominated strategies for another player, which will be eliminated as well. This process of successive elimination can continue until there are no more strictly dominated strategies left. If a unique strategy remains for each player, we say the game has been solved by *iterated dominance*. It is easy to prove that a strategy profile thus obtained is a Nash equilibrium (Bicchieri 1993).

Consider for example the game in figure 10.5 . *R* is a strictly dominated strategy for player 2, and since rationality is common knowledge, 2 is expected to eliminate *R* as a possible choice. Player 1 will now expect *L* to be played, in which case *U* dominates *D*. (*U*, *L*) is the solution to the game, and it is inferrable from CK1 and CK2. Note that assuming common knowledge of rationality (or at least some level of mutual knowledge of rationality) is crucial to obtaining the (*U*, *L*) solution. If there were some doubt about a player's rationality, the solution would unravel. For example, if 1 were to think there is a 0.01 chance that *R* is chosen, then he would be better off by choosing *D*. In real life this is likely to occur. That is, in real life a player may "play safe" and prudently choose *D*, but we are now discussing a completely different point. The question is not whether players are fully rational or believe each other to be. Rather, we want to know how far they can go in inferring a Nash solution from CK1 and CK2. As we have seen, the answer often is not far.

Predictability is hampered by another common problem encountered in game theory: multiple Nash equilibria. Suppose two players have to divide $100 among themselves. They must restrict their proposals to integers, and each must propose a way to split without knowing the other's proposal. If the total proposed by both is equal or less than $100, each gets what she proposed, otherwise they get nothing. This game has 101 Nash equilibria. Is there a way to predict which one will be chosen? Alternatively, is there a way a player can infer what the other will do, and thus adjust her proposal accordingly? In real life, many people would go for the 50/50 split. It is simple, and seems equitable. In Schelling's words, it is a *focal point* (Schelling 1960). A focal point equilibrium has some property that makes it salient. A solution may be salient because of historical precedent, or because it embodies cultural norms we share (Lewis 1969). Unfortunately, mere salience is not enough to provide a player with a reason for choice. In our example, only if it is common knowledge that the 50/50 split is the salient outcome does it become rational to propose $50. Game theory, however, filters out any social or cultural information regarding strategies, leaving players with the task of coordinating their actions on the sole basis of common knowledge of rationality (and of the structure of the game).

Consider now another game that many readers would intuitively know how to solve. The game of figure 10.6 has two Nash equilibria in pure strategies: (*C, C*) and (*D, D*), but in the (*D, D*) equilibrium each player plays a weakly dominated strategy. (*C, C*) is a *Pareto-dominant* equilibrium point, since it gives both players a higher payoff than any other equilibrium in the game. For this very reason, it should be a natural focal point for both players. This game is an example of a

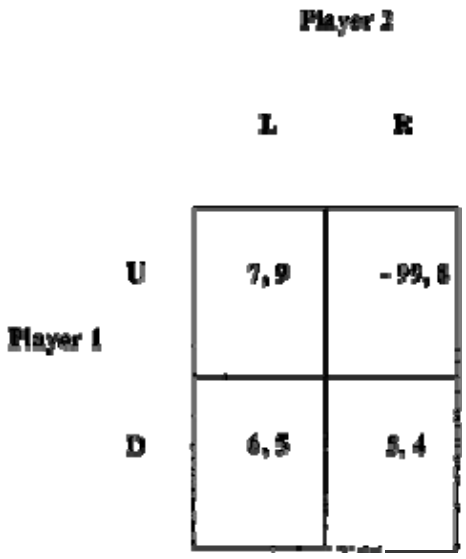

Figure 10.5

coordination game, that is, a game that has several Nash equilibria such that it is always preferable for a player to coordinate with the other and play one of the equilibria, since lack of coordination yields each player a lower payoff. Note that in the above example players' interests coincide, but even if we change the outcomes of (*C*, *C*) and (*D*, *D*) in figure 10.6 to (3, 1) and (1, 3), respectively, it remains true that both players prefer to coordinate upon one of the two equilibria to "going solo," even if each prefers a different equilibrium.

Should we confidently predict that (*C*, *C*) will be the solution of the game in figure 10.6? We have seen that focal points must be common knowledge among the players before it becomes rational for them to play the focal point equilibrium. If no such common knowledge is present, rationality alone is not a reliable guide. Could elimination of weakly dominated strategies do the trick? We know that

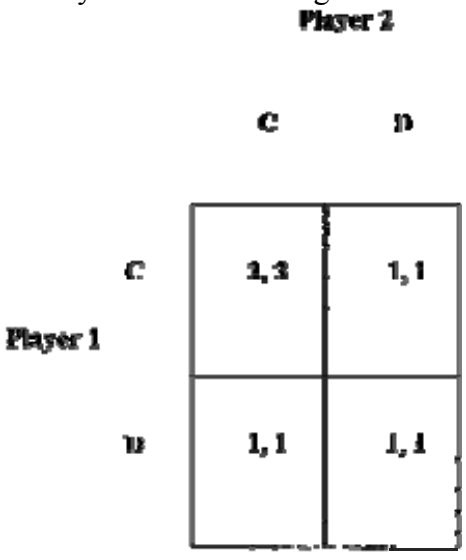

Figure 10.6
end p.193

when a player has a strictly dominated strategy, rationality dictates eliminating it, hence predicting behavior is a (relatively) simple matter. The case of weakly dominated strategies is not that straightforward. For one, a weakly dominated strategy is still a best reply to some opponent's strategy. Putting it differently, weak dominance means that there is at least one choice on the part of an opponent that makes one indifferent between the weakly dominated strategy and some other strategy. In our example, were player 2 to believe that 1 plays *D*, she would be indifferent between *C* and *D*, since both *C* and *D* are best replies to *D*, and conversely, were player 1 to expect 2 to play *D*, he would be indifferent between *C* and *D*, since both strategies are best replies to *D*.

One possible solution is to introduce a rule according to which weakly dominated strategies should also be eliminated by a rational player. The problem is that a player may be indifferent between two strategies, one of which is weakly dominated by the other, if she treats as null the state on which the weakly dominant strategy is strictly preferred. To guarantee that a player will always eliminate a weakly dominated strategy, we must assume that no state of the world is treated as null by the players. This means that a player's full belief that a strategy will be played is not to be interpreted as treating the event that it won't be played as null. As some philosophers have argued (Harper 1976 , McGee 1994), a state that is not considered an epistemically serious possibility need not be treated as null. This means that a player's conditional preference for a strategy, given such an “impossible” state, can be nontrivially defined. A related approach using lexicographic probabilities to reconcile iterated elimination of weakly dominated strategies and Bayesian decision theory is taken in Blume, Brandenburger, and Dekel 1991 and Stahl 1995 .

Another difficulty with applying iterated elimination of weakly dominated strategies is that the order of elimination usually matters; thus a particular solution will depend on which strategy was first eliminated. A standard solution to this problem is to delete at each round all weakly dominated strategies of all players (Rochet 1980 , Moulin 1986 , Harper 1991 , Bicchieri and Schulte 1997).

Eliminating weakly dominated strategies is an example of an “eductive” procedure. When asking how the players' deductive processes might unfold, one must specify some basic principles of rationality, and then examine which choices are consistent with common knowledge of the specified principles. Such choices may or may not result in an equilibrium, but at least the link between rational choice and equilibrium (when there is such link) is made clear. The advantage of this approach is that it is possible to refine our predictions about how players might choose without assuming that they will coordinate on a particular equilibrium. Principles such as iterated strict dominance and rationalizability are examples of how it is possible to restrict the set of predictions using rationality arguments alone. In most cases, however, the set of possible outcomes is still too large.
end p.194

A very different approach to the problem of indeterminacy is to start by considering the set of Nash equilibria, and to ask whether some of them should be eliminated because they are in some sense "unreasonable." This is the approach taken by the *refinement* program (Kohlberg 1990 ; van Damme 1991).

Normal Form Refinements

Consider again the game in figure 10.6. How reasonable is the equilibrium (*D, D*)? Under what circumstances would players agree to play it, and then stand by the agreement? The equilibrium strategies (*D, D*) are weakly dominated but—as I already argued—common knowledge of rationality does not force players to eliminate them. Prudence, however, may suggest that one should never be too sure of the opponents' choices. Even if players have agreed to play a given equilibrium, some uncertainty remains. If so, we should try to model this uncertainty in the game. Selten's insight was to treat perfect rationality as a limit case (Selten 1965). His "trembling hand" metaphor presupposes that deciding and acting are two separate processes, in that even if one decides to take a particular action, one may end up doing something else by mistake. An equilibrium strategy should be optimal not only against the opponents' strategies, but also against some very small probability > 0 that the opponents make "mistakes." Such an equilibrium is *trembling-hand perfect*. Is the equilibrium (*D, D*) perfect? If so, player 2's choice of *D* must be optimal against *C* being played (by player 1) with probability and *D* being played (by player 1) with probability 1− for some small > 0. But in this case player 2's payoff to *C* is 3 , whereas his payoff to *D* is . Hence for all > 0, *C* is a better strategy choice for player 2. Since the game is symmetrical, the same reasoning applies to player 1. The equilibrium (*D, D*) is not (trembling-hand) perfect, but (*C, C*) is. A prudent player therefore would discard (*D, D*). In this simple game, checking perfection is easy, since only one mistake is possible. With many strategies, however, there are many more possible mistakes to take into account. Similarly, with many players we may need to worry about who is likely to make a mistake.

Note that the starting point of this approach is the set of Nash equilibria of the game. It is assumed that players can calculate them and agree to play one. The goal now is to rule out all those Nash equilibria that are not reasonable agreements. In principle, an equilibrium that is reasonable under a given criterion of reasonableness might cease to be such under another, more restrictive criterion. Specifying why an equilibrium might be unacceptable is made easier by taking into account what *would* happen if one or more players were to "deviate" from

the agreed-upon solution. The reason is intuitive: a player should not agree to play his part in an equilibrium if—were the unexpected to happen—he would have been better off by playing another strategy. Therefore we may say that a crucial property required of an

equilibrium is that it is *stable* to players' deviating from it. To reason about “deviations” from equilibrium, it is helpful to have a richer description of the game. This is the reason why most of the refinement literature refers to games in extensive form, where the order in which players move and the information they have when making a choice are made explicit.

Games in Extensive Form

The *extensive form* of a game specifies the following information: a finite set of players $i = 1,n$, one of which might be nature (N); the order of moves; the players' choices at each move and what each player knows when she has to choose; the players' payoffs as a function of their moves; finally, moves by nature correspond to probability distributions over exogenous events. The order of play is represented by a game tree T, which is a finite set of partially ordered nodes tT that satisfy a precedence relation denoted by $<$. A *subgame* is a collection of branches of a game such that they start from the same node and the branches and the node together form a game tree by itself. In figure 10.7, for example, player 2's decision node as well as her moves form a subgame of the original game.
Whereas normal form games are represented by matrices, extensive form games are represented by trees. A matrix description shows the outcomes, represented in terms of players' payoffs, for every possible combination of strategies the players might choose. A tree representation is sequential, because it shows the order in which actions are taken by the players. It is quite natural to think of sequential-move games as being ones in which players choose their strategies one after the other, and of simultaneous-move games as ones in which players choose their strategies at the same time. What is important, however, is not the temporal order of events per se, but whether players know about other players' actions when they have to choose their own. In the normal form representation, players' information about other players' choices is not represented. This is the reason why a normal form game could represent any one of several extensive form games. When the order of play is irrelevant to a game's outcome, then restricting oneself to the normal form is justifiable. When the order of play is relevant, however, the extensive form must be specified.

In an extensive form game, the information a player has when she is choosing an action is explicitly represented using *information sets*, which partition the nodes of the tree. If an information set contains more than one node, the player who has to make a choice at that information set will be uncertain as to which node she is at. Not knowing at which node one is means that the player does not know which action was chosen by the preceding player. If a game contains information sets that are not singletons, the game is one of *imperfect information*. It may also be the case that a player does not remember what she previously did. In this case, the game is one of *imperfect recall*. All the games we consider here, however, will be ones of perfect recall, in that players will be assumed to remember what they did and knew previously.

A *strategy* for player *i* is a complete plan of action that specifies an action at every node at which it is *i*'s turn to move. Note that a strategy specifies actions even at nodes that will never be reached if that strategy is played. Consider the game in figure 10.7 . It is a finite game of perfect information in which player 1 moves first. If he chooses *D* at his first node, the game ends and player 1 nets a payoff of 1, whereas player 2 gets 0. But choosing *D* at the first node is only part of a strategy for player 1. For example, it can be part of a strategy that recommends "play *D* at your first node, and *x* at your last node." Another strategy may instead recommend playing *D* at his first node, and *y* at his last decision node. Though it may seem surprising that a strategy specifies actions even at nodes that will not be reached if that strategy is played, we must remember that a strategy is a full *contingent* plan of action. For example, the strategy *Dx* recommends playing *D* at the first node, thus effectively ending the game. It is important, however, to be able to have a plan of action in case *D* is not played. Player 1 may, after all, make a mistake and, because of 2's response, find himself called to play at his very last node. In that case, having a plan helps. Note that a strategy cannot be changed during the course of the game. Though a player may conjecture about several scenarios of moves and countermoves before playing the game, at the end of deliberation a strategy must be chosen and followed through the game.

The game of figure 10.7 has two Nash equilibria in pure strategies: (*Dx*, *d*) and (*Dy*, *d*). This is easy to verify by looking at figure 10.7, the normal form representation of the game. Is there a way to solve the indeterminacy? Representing the sequential version of the game as one of perfect information (figure 10.7) helps to solve it. Suppose player 1 were to reach his last node. Since he is by assumption rational, he will choose *x*, which guarantees him a payoff of 4. Knowing (by assumption) that 1 is rational, player 2—if she were to reach her decision node—would play *d*, since by playing *a* she would net a lower payoff. Finally, since (by assumption) player 1 knows that 2 is rational and that she knows that 1 is rational, he will choose *D* at his first decision node. The equilibrium (*Dy*, *d*) should therefore be ruled out, since it recommends an irrational move at the

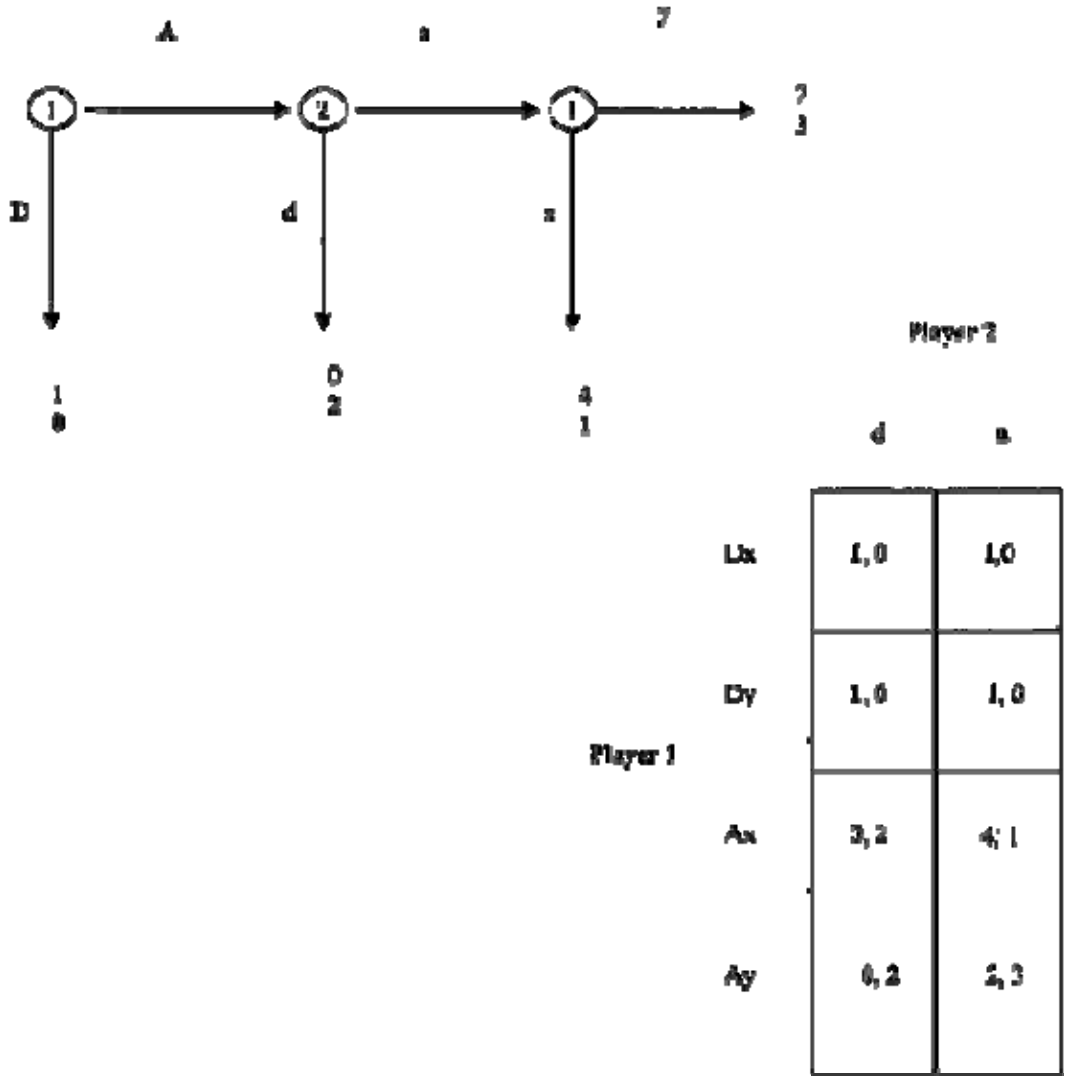

Figure 10.7

last node. In the normal form, both equilibria survive. The reason is simple: Nash equilibrium does not constrain behavior out of equilibrium. In our example, if 1 plans to choose *D* and 2 plans to choose *d*, it does not matter what player 1 would do at his last node, since that node will never be reached.

The sequential procedure we have used to conclude that only (*Dx*, *d*) is a reasonable solution is known as *backward induction* (Zermelo 1913). In finite games of perfect information with no ties in payoffs, backward induction always identifies a unique equilibrium. The premise of the backward induction argument is that mutual rationality and the structure of the game are common knowledge among the players (CK1 and CK2). It has been argued by Binmore (1987), Bicchieri (1989 , 1993), and Reny (1992) that under certain conditions common knowledge of rationality leads to inconsistencies. For example, if player 2 were to reach her decision node, would she keep thinking that player 1 is rational? How would she explain 1's move? If 1's move is inconsistent with CK2,

player 2 will be unable to predict future play; as a corollary, what constitutes an optimal choice at her node
end p.198

remains undefined. As a consequence of the above criticisms, the usual premises of backward induction arguments have come to be questioned (Pettit and Sugden 1989 , Basu 1990 , Bonanno 1991).
One should distinguish, however, the game theorist's informal justification of equilibrium play from players' reasoning to an equilibrium. A rigorous representation of agents' reasoning involves formally modeling their knowledge and beliefs, and encoding this information into a set of axioms. Players will then be represented as logical reasoners who infer (from a set of axioms and inference rules) a sequence of moves consistent with the axioms. A solution is precisely the sequence of moves players infer from their *theory of the game* (Bicchieri 1993). Part of the puzzlement with backward induction is due to the ascription of the informal backward induction argument to the players. Central to the informal argument is an analysis of out-of-equilibrium play, which is used to justify the choice of a given equilibrium (in this case, the backward induction one). If the backward induction argument is *formalized* as the players' own theory, it gives rise to an inconsistency when coupled with information that a deviation has occurred. The reason is obvious: any consistent theory that lets players infer an equilibrium becomes inconsistent when combined with a statement to the effect that a deviation from equilibrium occurred. It is only when considering deviations that inconsistencies may arise, but an analysis of deviations need not be part of the players' theory of the game. Such a theory may in fact harmlessly assume that players' rationality is common knowledge: The players would still succeed in computing the backward induction equilibrium (Bicchieri and Antonelli 1995).
Spurred by this problem, in recent years much foundational work has been devoted to carefully modeling players' knowledge and reasoning in games. Depending on how knowledge and hypothetical knowledge are represented, different authors have reached divergent conclusions about the link between backward induction and common knowledge of rationality. Some of the literature includes Bicchieri 1988b , 1993 ; Aumann 1995 ; Bacharach 1987 ; Stalnaker 1994 ; Samet 1996 ; and Arlo-Costa and Bicchieri 1998 .

Extensive Form Refinements

The goal of the refinement program, however, has not been the formalization of players' reasoning. The arguments proposed have been informal, their purpose being the elimination of implausible equilibria. In the normal form, Selten's trembling-hand perfection requires players to check how a strategy will perform were another player to take an action that has zero probability in equilibrium. In
end p.199

the extensive form representation, players ask what would happen off-equilibrium, at points in the game tree that will never be reached if the equilibrium is played. In both cases, the starting point is an equilibrium, which is checked for *stability* against possible deviations.

By its nature, the Nash equilibrium concept does not restrict action choices off the equilibrium path, because those choices do not affect the payoff of the player who moves there. For example, the equilibrium (*Dy*, *d*) in the game of figure 10.7 lets player 1 make an irrational choice at the last node, since that choice is not going to affect his payoff (which is determined by his choosing *D* at the beginning of the game). However, the strategy of a player at an off-equilibrium information set can affect what other players choose in equilibrium. Suppose the players consider agreeing to play (*Dy*, *d*). In order to choose *D*, player 1 must decide what would happen were he to play *A* instead. To decide whether *D* is a rational move, 1 has to think about player 2's choice at an off-equilibrium node. His conclusion about 2's choice will affect his own choice. But player 2's choice will depend upon how she interprets 1's off-equilibrium move. Player 1, in turn, must be able to anticipate 2's interpretation of his deviating from the equilibrium path. For example, if 2 were to interpret the deviation as a mistake, would she still play her part in the equilibrium (*Dy*, *d*), and choose *d*? If she expects *y* to be played at the last node, *a* is a best reply. But, at his last node, why would rational player 1 choose *y*? Is the agreement to play (*Dy*, *d*) reasonable?

The earliest refinement proposed to rule out implausible equilibria in extensive games of perfect information is *subgame-perfection* (Selten 1965). A Nash equilibrium is subgame-perfect if its component strategies (when restricted to any subgame) remain a Nash equilibrium of the subgame. The equilibrium (*Dy*, *d*) is not subgame-perfect: in the subgame starting at the last node, *y* is a dominated strategy. Note that the backward induction equilibrium is always subgame-perfect. Subgame perfection, however, applies only (nontrivially) to games that have proper subgames. Any Nash equilibrium of a game without proper subgames is trivially subgame-perfect (since the whole game can be considered a subgame), but in this case the criterion does not help in resolving the indeterminacy. In figure 10.8 , for example, both (*c*, *L*) and (*a*, *R*) are (trivially) subgame-perfect equilibria.

The game of figure 10.8 is a case in which we would still like to eliminate equilibria that require players to behave suboptimally in parts of the game that are reached with zero probability if a given equilibrium is played, but cannot be considered subgames. Kreps and Wilson's (1982) *sequential equilibrium* is an answer to this problem. A sequential equilibrium is a combination of strategies and beliefs such that each player has a belief (a probability assessment) over the nodes at each of his information sets. At any information set *x* where *i* has to play—given player *i*'s beliefs at *x* and the equilibrium strategies of the other players—*i*'s strategy for the rest of the game must still maximize his expected payoff. As

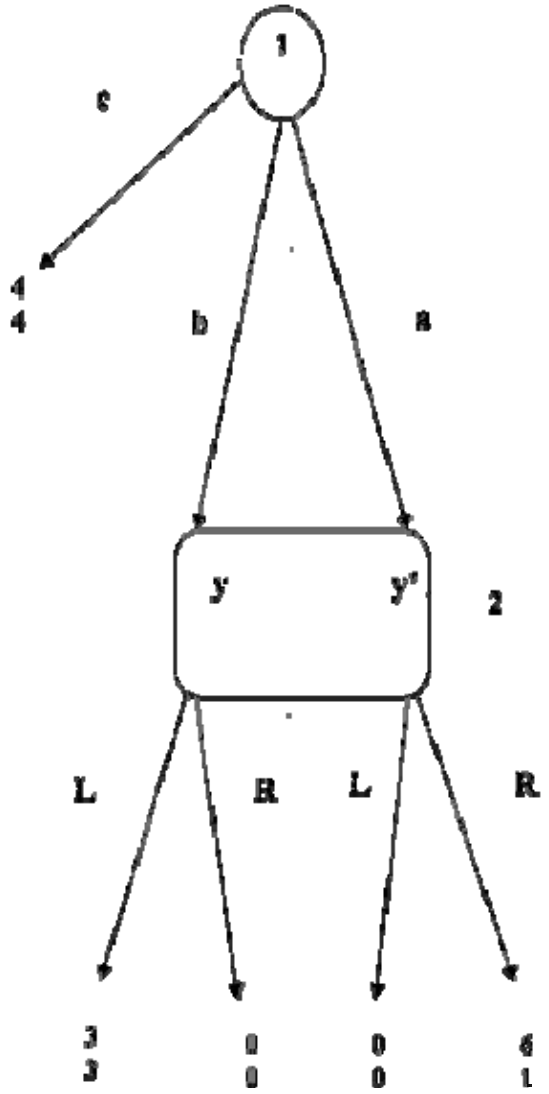

Figure 10.8

players move through the game tree, they rationally update their beliefs using Bayes's rule. The problem with the notion of sequential equilibrium is that—provided beliefs are revised according to Bayes's rule—no further restriction is imposed upon them. The consequence is that far too many Nash equilibria are still considered admissible or "reasonable." At player 2's information set, if for some reason she assigns a higher probability to node *y* than *y*', then her optimal choice is *L*. If instead she judges *y* and *y*' to be equiprobable, she will choose *R*. Thus both (*c*, *L*) and (*a*, *R*) survive as sequential equilibria, since each is supported by some acceptable belief.

Another common refinement is Selten's *perfect equilibrium* (Selten 1975). In this case players are explicitly assumed to interpret deviations from equilibrium play as "mistakes" and respond accordingly. A perfect equilibrium must be robust to small perturbations of players' equilibrium strategies. Selten's notion, however, by not imposing restrictions upon players' beliefs, lays itself open to the same criticism addressed to Kreps and

Wilson's refinement. If there are several possible "mistakes" a player can make, and beliefs are unrestricted, some equilibria cannot be ruled out simply because they are supported by beliefs that make some mistakes more likely than others. Suppose that in our example player 2 believes that player 1 intends to play c with probability $1–p–q$ very close to one, but can play b by mistake with higher probability (p = 2/100) than playing a by mistake (q=1/100). If this is what 2 believes, she should choose L. Given her beliefs, L has an expected utility of .06, whereas R has an expected utility of .01. Both equilibria therefore survive some perturbations.

It may seem that introducing a restriction as to which mistakes are more likely to occur would help in limiting the number of plausible Nash equilibria. Myerson's *proper equilibrium*, for example, is a Nash equilibrium that is robust with respect to "plausible" deviations only, by which he meant deviations that do not involve *costly* mistakes (Myerson 1978). Yet what makes a mistake costly is not a simple matter, since it depends upon what players believe other players will do in reaction to a mistake. Consider again the game in figure 10.8. If a deviation from (c, L) were to occur, player 2 would keep playing L only if she believed the probability of mistake b to be greater than that of mistake a. And if player 1 were to believe that 2 would respond to a deviation with L, mistake b would indeed be less costly for him than mistake a. In this case, again, both equilibria will survive.

Note, however, that strategy b is strictly dominated by c for player 1, therefore it is highly unlikely that 1 would play b. This example is meant to stress the importance of restricting off-equilibrium beliefs. Such beliefs, too, should be rationally justified. The equilibrium (c, L) must be eliminated because it is supported by the belief that a dominated strategy will be played off-equilibrium, and this belief is inconsistent with common knowledge of rationality.

The need to add a condition of plausibility for off-equilibrium beliefs motivated the *forward induction* refinement (Kohlberg and Mertens 1986). Off-equilibrium beliefs, for example, should be consistent with common knowledge of rationality and any inference one may draw from it. A deviation from equilibrium should therefore be interpreted, whenever possible, as a rational move. In our example, player 1's deviation from the equilibrium (c, L) should not be interpreted as a mistake, but rather as a *signal* that he intends to play a (and get a higher payoff). In this case player 2 would respond with R. The conclusion is that equilibrium (c, L) is not robust to deviations, and should be eliminated.

As I mentioned at the outset, the refinement program attempts to establish stability criteria for Nash equilibrium. It presupposes that players will choose to play a Nash equilibrium after having eliminated several alternative equilibria on the ground that they are unreasonable. What counts as a reasonable equilibrium,

however, depends upon how off-equilibrium behavior is interpreted. This, in turn, hinges on players' out-of-equilibrium beliefs. A formalization of the off-equilibrium deliberation process requires the use of counterfactuals (from the viewpoint of playing a given equilibrium, deviations are contrary-to-fact events). Some work in this direction has been done by Selten and Leopold (1982), Bicchieri (1988a), Harper (1991), Samet (1996),

Stalnaker (1994 , 2000), and Skyrms (1990 , 1998). So far we have developed no comprehensive theory of out-of-equilibrium behavior that indicates, for example, when a deviation should be interpreted as a signal and when as a mistake. Such theory would supply substantive (as opposed to merely formal) rationality criteria for players' beliefs, and would thus expand the traditional notion of practical rationality to include an epistemic component. This theoretical inadequacy undermines the eductive goal of inferring (and predicting) equilibrium play from rationality principles alone.

Repeated Games

The games we have discussed so far were all one-shot games. Many real-life interactions, however, are ongoing ones. Interesting social phenomena such as cooperation, retribution, trusting and reciprocating, committing and promising can be represented by means of repeated games. Take the Prisoners' Dilemma (see also Dreier, chap. 9, and Sorensen, chap. 14, this volume) depicted in figure 10.1. The label *C* now stands for "cooperate" and *D* for "defect." [1] What should a rational player do? Notice that the payoffs represent ordinal preferences; thus by just looking at the matrix it is clear that each player prefers to defect, whatever the other does. In fact, to defect is a dominant strategy for both. But notice that, if both were to cooperate, their payoffs would be superior to those of joint defection. Mutual cooperation is Pareto-optimal, but it is not an equilibrium. Of course the players may have different preferences, and like cooperation better than defection, but then we would have to draw a different matrix. In the game of figure 10.1, players' preferences doom them to an unsatisfactory outcome.
Suppose now that the Prisoners' Dilemma is repeated *N*-times between the same two players. Since the game is noncooperative, the players cannot enter into binding agreements, so whatever threat or promise they make to each other, it must be enforceable to be credible. Suppose that it is common knowledge among the players that the game is ending at the *N*th repetition. In this case, both know that in the last repetition they will both defect, irrespective of what they said or
end p.203

did before. At the *N–1*th stage then, they will both know what will happen at the *N*th stage, so the rational choice will be to defect there, too. Proceeding inductively, both players will reach the conclusion that each will (and should) defect at each stage. The players use backward induction (see also Sorensen, chap. 14, this volume) to reach this conclusion, and the only Nash equilibrium of this repeated game is to defect throughout. Since the payoffs to each player are the sum of what she receives at each stage game, the players will end up with *1N* each, whereas joint cooperation would have netted them *3N* each. I am assuming here that players have common knowledge of the game, payoffs, and mutual rationality. Relaxing any of these assumptions will give us different results, one of which is that players can rationally sustain cooperation for most of the game (Kreps and Wilson 1982 ; Bicchieri 1993).

Keeping firm the traditional assumptions, however, we can still achieve cooperation in a repeated Prisoners' Dilemma, provided the game is infinitely repeated or, alternatively, there is an unknown termination and a sufficiently high probability that the game continues another stage. To see why cooperation can be achieved, suppose that, after each repetition, there is a probability of 1/3 that this is the last stage. Then the probability that the game will persist until at least the *N*th stage is $(2/3)^{N-1}$. Note that the probability that the game is eternal is zero, since $(2/3)^{N} \rightarrow 0$ as $N \rightarrow \infty$. Now consider the following strategy *S*: "Play *C* as long as the other plays *C*, but if he plays *D*, play *D* forever after." If both players adopt *S*, they will cooperate forever. Player *i*'s expected payoff of playing *C* at each stage is $3 + 3(2/3) ++ 3(2/3)^{N-1} + 3(2/3)^{N} + 3(2/3)^{N+1} +$.Can a player gain by deviating? Suppose player *i* plays *D* at the *N*th stage. If the other player plays *S*, the deviant player will at most get the following payoff: $3 + 3(2/3) ++ 3(2/3)^{N-1} + 4(2/3)^{N} + 1(2/3)^{N+1} +$.For a deviation to be unprofitable, it must be the case that the expected payoff of continuous cooperation (enforced by *S*) is greater or equal to the expected payoff of, say, deviating at the Nth stage. In our example, $C - D = (3–4)\ (2/3)^{N} + (3–1)\ (2/3)^{N+1} + (3–1)\ (2/3)^{N+2} +$, which amounts to $(2/3)^{N}\ (-1+4)$. Since this value is greater than zero, this means that the expected payoff of continuous cooperation is better than the expected payoff of deviating at the *N*th stage. Therefore (*S*, *S*) is a Nash equilibrium. In fact, we know that in our game any feasible expected payoff combination that gives each player more than 1 can be sustained in equilibrium. This result is known as the *Folk Theorem*, so called because it is uncertain who proved it first but it has become part of the folk literature of game theory.

What is rational in a one-shot interaction may not be rational when the game is repeated, provided certain conditions obtain. The conclusion we may draw is that phenomena such as cooperation, reciprocation, honoring commitments, and keeping one's promises can be explained as the outcome of rational, self-

end p.204

interested choices. In repeated interactions, even a person who cares only about material, selfish incentives can see that it is in his interest to act in a cooperative way. Whether this is a good, realistic explanation of prosocial behavior is another story.

NOTE

1. The most familiar version of the Prisoners' Dilemma is found in Luce and Raiffa 1957 , 94–97.

end p.205

chapter 11

PRACTICAL REASONING AND EMOTION

Patricia Greenspan

The category of emotions covers a disputed territory, but clear examples include fear, anger, joy, pride, sadness, disgust, shame, contempt, and the like. Such states are commonly thought of as antithetical to reason, disorienting and distorting practical thought. However, there is also a sense in which emotions are factors in practical

reasoning, understood broadly as reasoning that issues in action. At the very least emotions can function as "enabling" causes of rational decision making (despite the many cases in which they are *dis*abling) insofar as they direct attention toward certain objects of thought and away from others. They serve to heighten memory and to limit the set of salient practical options to a manageable set, suitable for "quick and dirty" decision making.

Current research in neuroscience and other areas indicates that practical reasoning in this sense presupposes normal emotional development and functioning (see, e.g., Damasio 1994). Evolutionary accounts of emotions (e.g., Cosmides and Tooby 1992) stress their role in the rational design of the human organism. Contemporary philosophy of emotion attempts something stronger, however, in according emotions a role in practical reasoning. Making this an integral role—understanding emotions as functioning within practical reasoning rather than just as spurs *to* it—means interpreting emotions in normative terms, as providing or expressing potential reasons for action, and as themselves subject to rational assessment and control, contrary to the traditional view of emotions as "passive" phenomena.

1. Emotions as Evaluative

The dominant approach in contemporary philosophy rests on assigning emotions an evaluative content (see, e.g., Bedford 1957 , Solomon 1976 , Lyons 1980 , Budd 1985 , Davis 1987 , de Sousa 1987 , Roberts 1988 , Greenspan 1988 , Nussbaum 1993 , Stocker 1996). Human emotions typically are directed toward "intentional objects" in the sense of being about something—real or imagined, but in any case an object of thought. Emotions that represent their objects in some positive or negative light (as most do) may be said to have a content expressible by an evaluative proposition. Where the evaluation either is or implies an evaluation of some future contingency that the agent (the one who undergoes the emotion) can bring about or avoid, the emotion provides a reason for or against action.

The rational bearing of emotions on action in these terms can be captured in simple form by the first or major premise of an Aristotelian practical syllogism, Aristotle's three-line schema of practical reasoning (Aristotle 1984 : *Nicomachean Ethics*, bk. 7; for an alternative model, see Thagard 2000). In contemporary terms the major premise expresses a pro- or con-attitude toward something that the second or minor premise tells one how to attain or avoid. It evaluates something as good or bad, desirable or worth avoiding—or simply as an object of a current desire or aversion.

For instance, fear may be said to represent its object (what one is fearful of or about: a certain event, or a likely cause of it) as a threat, a possible harm. Anger represents its object (another person, or his performance of some action) as already a cause of harm or offense that now calls for retaliation. Despite differences between the propositional content of these two emotions—one describes a possibility, the other a fact (Gordon 1987)—both include a reason for avoiding some future contingency: the presence or occurrence of what is feared or a failure to retaliate for the cause of anger.

Emotions also involve a corresponding affective element, derived from pleasure or pain on Aristotle's account of emotions (1984 : *Rhetoric*), that amounts to a good or bad state of the agent and hence supplies a reinforcing reason for action (Greenspan 1988). In contemporary decision-theoretical terms, the affective element modifies the "payoff structure" of the situation: the array of potential costs and benefits of alternative responses open to the agent. The discomfort experienced in fear or anger, for instance, provides the agent with a further (*pro tanto*) reason for acting to change the situation that provokes discomfort. On the other hand, joy or pride, as positive evaluations of some state of affairs or of oneself, do not provide a reason to change anything, but their affective aspect yields a further reason to sustain the conditions that make the evaluation appropriate. However, the focus of the agent's attention in all these cases is normally the evaluative content of emotion, not her own state of feeling.
end p.207

Apart from its role as a source of reinforcing reasons, the affective or feeling aspect of emotion is useful to practical reasoning just insofar as it serves to hold in mind the evaluative content of emotion without explicit reflection. For instance, while driving on the highway one does not have to deliberate at length about possible bad consequences of swerving and their relevance to the task of steering straight. To the extent that steering straight is not just automatic, and the driver wanders out of her lane, an anxious awareness of the possibility of an accident brings her quickly back to the task.
This way of anticipating practical eventualities in everyday life corresponds roughly to neuroanatomist Antonio Damasio's (1994) understanding of emotions as "somatic markers." Damasio's view is put forth to explain why cases of emotional impairment due to brain lesion (such as the famous nineteenth-century case of Phineas Gage) involve a loss of practical reasoning ability. Emotions serve to "mark" practically significant thoughts with bodily (and hence affective) indicators of past experience. According to an evaluative account, characteristic thoughts have come to be contents of emotion—and part of what identifies them as the types of emotion they are: fear, anger, joy, pride, and so forth.
A question for philosophers who accept an evaluative account is whether and in what sense emotions themselves are subject to rational assessment. It is only in a qualified sense that an irrational evaluative stance could be said to justify action. Hume's famous denial (see 1978, bk. 2, pt. 2, sec. 3) that the passions can be rationally assessed at all was based on his interpretation of them as nonrepresentational. But contemporary evaluative accounts interpret emotions, in effect, as representing evaluative propositions: *that* something portends harm or deserves retaliation or is a stroke of good luck or praiseworthy.

2. Belief-Based Views

The evaluative implications of emotion have been characterized thus far without commitment to a particular contender for capturing their essential nature. This is the

philosopher's usual first question, and it was set up as the focus of contemporary debate on emotion by James (1884). However, as Rorty (1980c) and others have noted (cf. Griffiths 1997), emotions do not form a natural class. A later approach (de Sousa 1987 , Greenspan 1988) begins with questions of rationality, treating cases of rational emotional response as paradigmatic. Emotions in lower animals or infants (along with some elements of our own experience that
end p.208

derive from them unchanged) can be understood from this standpoint as deficient instances of full-blown emotions functioning normally in adult human life.

As applied to accounts of the nature of emotions, the evaluative approach is often discussed under the heading of "cognitivism," which interprets emotions themselves as amounting to or containing cognitions, usually seen as mental states representing evaluative propositions. The most straightforward version of the cognitivist view is "judgmentalism," which understands the nature of emotions in terms of judgments: assertive propositional attitudes, assessable by the ordinary evidential standards applied to beliefs. The view traces back to the Stoics (see Nussbaum 1993), though the Stoics interpreted our ordinary emotions as "confused" rather than rational judgments.

The contemporary version of judgmentalism was introduced into the Anglo-American philosophical literature by Bedford (1957), in a Wittgensteinian response to the Cartesian tradition of identifying emotions with particular passions or feelings, as in Hume (cf. Descartes 1984a , James 1884). The view was extended by Kenny (1963) and Pitcher (1965), and it was brought into connection with other philosophical traditions and with work in psychology and other areas by Solomon (1976), Neu (1977), Lyons (1980), and Ben-Ze'ev (2000). Davidson (1980) defends a version of judgmentalism for pride and related emotions.

Proponents of judgmentalist accounts do not always seem to have the same thing in mind by "judgment." Diverse cases indicate, moreover, that the view needs some sort of qualification to capture situations where emotions register evaluative thoughts that may not rest on evidence supporting judgments in the sense of all-things-considered beliefs but nonetheless seem to be rational to hold in mind and to act in light of.

Some irrational cases such as phobias (fear of a dog one knows to be harmless, say) that undermine judgmentalism as a view of the nature of emotions are explored at length in Pitcher 1965 , Stocker 1987 , and Greenspan 1981 and 1988 . However, there are also many morally or motivationally important emotions that one might be inclined to assess as rational but that turn on imagining some state of affairs that is not applicable (at least in current terms) to oneself.

Cases of empathetic emotion, for instance, seem to involve putting oneself in the position of another person, typically someone suffering harm. However, acting effectively out of sorrow felt on behalf of someone else often depends on not losing sight of the fact that one is not really in the same position oneself. Similarly for anticipatory emotions such as guilt at the thought of something one might (or might not) do. There are also "emotional self-management" cases—of generating upbeat feelings, say, by "looking at the bright side" of some more problematic object of evaluation, or looking past someone's flaws to sustain love—that are crucial to mental health and to thriving interpersonal relationships.

end p.209

3. Modifying Judgmentalism for Rationality Issues

One might distinguish a possible variant of judgmentalism meant to apply specifically to rationality issues. Whatever their nature as psychological states, according to this account, emotions can be understood in terms of some sort of claim they make about the agent's situation. Whether a given emotion is rationally appropriate depends on whether this corresponding judgment would be *warranted as* a belief—whether or not the emotion actually involves that belief.

Emotions that do not conform to this model, such as empathetic or anticipatory emotions, might still sometimes be said to play a limited instrumental role in practical reasoning as cases in which an irrational mental state serves as part of an overall rational strategy. This idea of emotional motivation as "rational irrationality" comes up particularly in discussion of evolutionary design strategies that might account for emotional proclivities underlying human morality (see Frank 1988).

Thus, one could grant that feeling sympathetic sorrow over someone else's loss would not entail a belief that one has suffered some sort of loss oneself, or warrant for such a belief. But on the present suggestion that is just because it is not really appropriate to one's own situation. The feeling may still be important to sustaining mutually supportive personal relationships, and we can encourage it in ourselves with that end in mind, though to preserve its other-directed quality, we usually try to keep such calculations subliminal, or refer them to evolutionary design.

Whether this is a satisfying answer for all cases depends in part on how we think of rationality. In common parlance the term is sometimes used in a "thick" sense that would be implausible for most emotions insofar as it connotes explicit calculation; by contrast, Gibbard 1991 uses the term in an extremely "thin" sense in which it conveys just general positive endorsement. But there seems to be room between these poles for a use that captures distinctions we commonly make between reasonable and unreasonable emotional reactions, appropriate and inappropriate in what seems to be roughly a rational sense, having to do with some notion of fit to the circumstances that constitute grounds or evidence for emotion. We might think of this medium-thick sense as "representational" rationality insofar as it depends on assigning emotions intentional content.

However, our ordinary notion of emotional appropriateness seems to allow for variable standards of evidence and for more than one option in many cases. For instance, under ordinary circumstances it might be reasonable enough to feel somewhat worried about how a recent medical test will turn out, even in the absence of grounds for real agitation. At the same time, there may be no rational

end p.210

barrier to foregoing any thought about the matter or even feeling hopeful or confident. Being seized with fear would be irrational, let us say, not just because it would do no

good, but more fundamentally because there are no grounds for holding in mind with that intensity any subset of the available evidence that would warrant the belief that one is in danger.

An emotion might come out as rationally appropriate, according to this way of looking at things, even in cases where the corresponding all-things-considered belief would not be warranted. In other words, emotional appropriateness does not seem to be determined by a unitary overall assessment of the available evidence, of the sort presupposed by belief warrant. There is at most one adequately warranted belief (over the threshold for assent) about the probability of a bad outcome, relative to the evidence available to the agent. Belief may admit degrees, but where the evidence is slight, it would be irrational to adopt even a low-grade belief, if that entails adopting conflicting beliefs. However, the emotional analogue (ambivalent feelings, in some cases even intense feelings) would seem to be rationally acceptable—rational in a relatively undemanding sense that allows for options (Greenspan 1980).

We can think of this as "perspectival" rationality. Moderate worry and hope in the case just outlined may each be rational from different perspectives—with reference to different subsets of the available evidence, each worthy of the relevant degree of emotional attention, though not requiring it. A perspectival account of emotional rationality would appeal to a notion of warrant for a prima facie belief, as determined from some practically significant standpoint of evaluation (Greenspan 1988 and 1995). So it still represents a fairly conservative move away from judgmentalism on rationality issues—though it does move beyond the view just distinguished, which makes out emotional appropriateness as dependent on warrant for a corresponding belief. The perspectival account associates emotions with evaluative thoughts rather than strictly judgments in the sense of beliefs (or acts of forming beliefs), but it still allows that the rationality of emotions can be explained in terms of belief warrant.

4. A Causal/Historical Approach

In a more radical departure from judgmentalism, emotions might be thought to be justified rationally by reference to their causal histories rather than some notion of evidence of the sort that applies to beliefs. In light of the prominence of causal theories in other areas of philosophy, a causal/historical approach to emotions would seem to provide an alternative to judgmentalism and its variants. Someapproach along these lines is suggested by recent evolutionary treatments of emotion in philosophy (e.g., Gibbard 1991, Griffiths 1997). However, the only example of it that has been formulated to apply to issues of rationality in the sense indicated is the "paradigm scenarios" view in de Sousa (1987), which is still phrased in roughly cognitivist terms.

According to de Sousa's account, whose initial sources are psychoanalytical, emotions are based on infantile response tendencies such as smiling or crying that become intentional or object-directed and thereby come to be signs or elements of emotions by association with a typical story about what prompts the response. The story is what de Sousa calls a "paradigm scenario," meaning that it sets the standard for a particular emotion type, essentially by establishing criteria of appropriateness. Originally, what is involved here is the child's interaction with caretakers, which serves to modify natural responses by a form of dramatic role-playing (eliciting or exchanging smiles, say, or conveying empathy). Later cultural influences can modify the emotion, though it will remain appropriate in some form as a response to the paradigm scenario.

A central example de Sousa wants to defend as appropriate in these terms, in response to some feminist arguments, is jealousy, or jealous rage, with psychoanalytic sources shaped by monogamous culture. According to his account the emotion is set up as a response to a situation of "being robbed by another of vital physical attention" (185). That is how it gets its status and meaning as an intentional response, so it necessarily counts as appropriate on that basis.

It is unclear, though, whether the paradigm scenarios account can explain more than the appropriateness of (at least some central cases of) a given *type* of emotion. De Sousa's essential point about rationality in the sense of appropriateness seems to be just that there cannot be an emotion type of which all possible instances are inappropriate. However, it does not follow from this that a given emotion instance should be assessed as appropriate on the basis of its resemblance to the paradigm scenario for its type. More specific rationality questions—under what circumstances jealousy is appropriate in adult life, say—remain open.

De Sousa speaks of the rationality of emotions in several senses, but overall he seems to be most concerned with rationality in the sense of objectivity. On that interpretation, emotions are rational in general terms to the extent that they represent real properties of the world, existing independently of emotions. This amounts to an emotion-based version of the metaethical question of the objectivity of values, since the properties in question, corresponding to different types of emotion, are "axiological" or value properties. However, recent views within the metaethical literature would seem to allow for emotional rationality in a sense independent of objectivism about values (see, e.g., Gibbard 1990 , Blackburn 1998 ; cf. Greenspan 1995 and 1998 , D'Arms and Jacobson 2000). Objectivity need not be our concern when we speak of the rationality of particular emotions in the sense of appropriateness, but rather their

fit to the reasons available to the agent, leaving it as a further question whether reasons or values in general can be thought of as objective.

De Sousa at some points limits himself to a claim of "minimal rationality" or basic intelligibility of emotion types, but he also applies the paradigm scenarios view to the question of the correctness or incorrectness of particular emotion instances, which his discussion makes out as dependent on the notion of normality (cf., e.g., 187, 201). But some emotional responses that are normal and understandable might be inappropriate. We can make sense of a claim, for instance, that the paradigm for male jealousy—what degree and sort of attention the jealous agent feels he is entitled to and hence can be "robbed" of—has been distorted by male upbringing in a certain culture. So even the received paradigm of an emotion can be skewed or off the mark.

Applying the paradigm scenarios approach to cases would seem to involve appeal to an interpretation of the situation that might be summed up in a proposition: that one has been robbed of vital physical attention, in the jealousy example. How this gets extended or modified by culture (extended to other forms of attention, perhaps, or modified by challenging the criteria of entitlement to attention) may be a matter of historical derivation, but the criteria at a given time and relative to a given culture would seem to be specifiable in terms that fit the belief-based approach outlined earlier. De Sousa rejects

the analogy between emotions and beliefs as a basis for his approach to rationality issues at least partly because he understands rational beliefs as arrived at in a certain way, on the basis of reasoning (see 5; cf. 197ff.). However, this is not part of the belief-based approach as here intended.

It is not clear how the paradigm scenarios view could handle cases without appeal to belief-based criteria of rationality. For example, imagine someone who momentarily feels jealous anger when his wife exchanges glances with another male at a party. In de Sousa's terms the man is reacting to what he sees as a robbery of vital attention—in terms of Aristotle's definition of anger (see 1984 : *Rhetoric*, 378a31–b5), an unjustified slight, or at any rate an indication that a slight is imminent. But this is an evaluative judgment interpreting past events and their natural and conventional meanings: what a glance means or can mean, what legitimate expectations a relationship confers, that a glance involves or might lead to intimacies that violate those expectations, and that violating them inflicts loss and insult on the agent.

The emotional reaction in such a case might sometimes be assessed as inappropriate to the current situation—if the agent really knows, say, that his wife and the recipient of her glance, a colleague in her area, are merely reacting to a professional faux pas on the part of someone else at the party—even where situational cues do naturally give rise to jealousy because of their superficial resemblance to the paradigm scenario. To explain which cues render an instance of emotion appropriate, rather than merely natural or normal or understandable, we

seem to need at least implicit reference to the notion of a propositional content, as what the emotion still essentially “claims” about the situation, even if the reasons for it have changed since the paradigm scenario was established.

5. Emotional Strategies

A variant of the general causal-historical approach that rejects reference to propositional attitudes is the evolutionary account in Griffiths 1997 . Though Griffiths is mainly concerned with the question of the nature of emotions rather than rationality issues, his argument turns out to rest at crucial points on a familiar way of setting up emotions in opposition to reason. He continually adverts to a definition of emotions as “irruptive” or passive states interfering with the agent's stable goals and not themselves subject to rational strategy (see, e.g., 155–57, 233ff., 242ff.; cf. 9, 16, 118, 120). Those cases in which emotions appear to be used strategically cannot be genuine emotions according to his account.

Griffiths particularly has in mind emotions based on the social pretense that we lack rational control over a certain pattern of behavior. He draws on anthropological examples from defenders of “social constructivist” views of emotions (see, e.g., Averill 1980 , Lutz 1986) in which the conditions of particular syndromes of emotions seem to be based on social conventions. A parallel from contemporary urban life is the “Rambo” syndrome of

violent male rage; some conceptions of romantic love might also qualify. However, Griffiths's point is that whatever is felt in such cases would be disqualified from inclusion in a class of genuine emotions.

However, it seems, on the contrary, that we can make sense of a *self-fulfilling* social pretense of emotion in the example of male jealousy drawn from de Sousa. Imagine someone working himself into states of jealous anger on flimsy or imagined grounds (or even on good grounds that he normally would ignore) in order to provoke a certain kind of interaction with his spouse—to exert control, perhaps, or perhaps just to bring about an occasion to express and enhance affection.

An attempt to understand emotions as voluntary actions, things we do rather than states that come over us, was part of Solomon's (1973) version of judgmentalism, based on Sartre's (1962) conception of emotions as "magical transformations of reality" but meant to allow for rationality. However, we need a less extreme version of the view to modify the usual picture of rationality and emotions. The rational role of emotions depends on the fact that our control over them is limited.

Thus, for instance, Frank (1988) assigns emotions a crucial role in answer to

what he calls "the commitment problem," the problem of how we can demonstrate to others effectively our commitment to future actions against narrow self-interest, as needed to secure their willingness to take part in cooperative schemes (on the model of the Prisoners' Dilemma, where each agent has to resist inducements to confess a joint crime). Emotions are as useful as they are to social communication just by virtue of the fact that they are not states an agent can simply produce at will.

So the question is whether we can strategically bring about genuine states of partial uncontrol, individually or collectively. But the answer seems to be that we do this all the time. I would add to Frank's account the suggestion that we can do it without sacrificing rationality—even in the short term, to the extent that we undergo an emotion. The point depends on understanding emotions themselves as potentially rational and as rationalizing action, but also on appreciating features of their role in directing attention that make them subject to manipulation by the agent (cf. Mele 1995).

An effective emotional strategy has to be somewhat indirect: we cannot just "talk ourselves into" sincere emotional states by commanding ourselves to feel them or by taking immediate aim at the signs of emotion. Instead, we set up the conditions under which they would arise. These include conditions of thought as ways of understanding the situation.

Imagine working up an emotion in order to argue effectively. A case I used in Greenspan 2000 involves letting oneself get angry about a consumer complaint when the time comes to confront someone at the store. But one might also consider the everyday occasions in teaching or giving a paper when one works up a strategic emotion: indignation at opposing views or enthusiasm for a question that is not of current concern. This might be (and remain) just a pretense, but it also might be a real result of focusing attention on aspects of the material and the surrounding situation that are likely to generate the requisite emotion: the wrongheadedness of the opposing view or the features of one's own view that initially made it seem worth defending.

Some people are better at this than others, needless to say. Part of what it takes is control over attention—not entertaining those self-defeating thoughts one may well have on hand, of sympathy for the opposing view and despair at the imperfections of one's own. Also, it is crucial not to focus too directly on strategy: the aim of generating an emotion in oneself that will enhance performance.

To count this restriction of attention as rational in decision-theoretical terms, we have to be clear that rationality does not require being immediately and centrally aware of everything one knows. According to standard assumptions in decision theory it does require full *knowledge* pertaining to the choice situation. But putting blinders on oneself in this sense (or passively accepting them) is "within reason," as long as what one knows remains within reach, perhaps on the periphery of attention.

end p.215

A perspectival evaluative account of emotions and emotional rationality can accommodate such maneuvers. We have two competing partial views of the situation—one focusing on positive features, one on negative—and some choice about where to focus. So for different purposes we can shift the focus, thereby generating the relevant emotional state—not with any certainty but with enough probability for the maneuver to count as an exercise of rational control. Since the state involves positive or negative affect (or possibly a mixture of both), we have essentially modified the payoff structure by adding emotion to the preexisting situation.

This account is compatible with the usefulness of emotions as commitment devices, which requires that our control over them be incomplete. For purposes of commitment it is enough that emotions be difficult to fake convincingly and that they make certain courses of action (or refraining from action) difficult. This means that emotional states, or those that are useful as commitment devices, have to exhibit a certain momentum: we cannot so easily just talk ourselves out of them, including those we managed to talk ourselves into.

In the case of generating enthusiasm to give a paper, momentum (with any luck) will be a product of the rewarding aspect of the feeling, as augmented by others' positive responses. But in a commitment case like Prisoners' Dilemma something more is possible: if the prisoners have sworn loyalty to each other and firmed up their feelings accordingly, that means instilling negative reactions to the thought of betrayal that will be hard to go against when the opportunity to gain by confessing arises.

This case is in some ways easier in that it involves inducing dispositions to feel certain emotions rather than occurrent emotions. The notion of cultivating certain traits in ourselves, including emotional dispositions, is familiar from Aristotle's account of habituation in virtue (1984 : *Nicomachean Ethics*) and still plays a role in philosophers' views of emotional motivation (see, e.g., Wollheim 1991). However, most recent approaches to emotion focus on occurrent states, perhaps in reaction to an earlier overemphasis on dispositions in philosophical behaviorism (see Ryle 1949). To appreciate the rational role of emotion, we need to consider uses of occurrent emotions such as anger (to back up a threat to call the Better Business Bureau, in my consumer case), as well as the sorts of commitments an agent can make to himself—to follow through on a resolution (possibly by "holding onto" his feelings of anger).

Damasio's point about “marking” thoughts with emotion covers some cases of emotional self-commitment. Consider learning a foreign language: one way of implanting information in long-term memory is to focus emotional attention on it—perhaps by thinking of some association that evokes amusement. Examples involving practical reasoning include viewing the prospect of going back on one's resolutions—analogous to betrayal in Prisoners' Dilemma—with self-disgust or some other version of anticipatory guilt, bringing on a motivating emotion now as well as generating a disposition to feel it later.

6. Emotions as Reasons

By inducing an occurrent emotion, I would say, we give ourselves a further reason for action—much as we do when we make a promise to others. We can count an emotion as itself a reason in the sense of a motivating reason, a state of thought, or as a source of reasons in a more objective sense, of normative reasons as facts about what needs to be done to sustain or ameliorate a state of emotional comfort or discomfort. However, this rational role of emotions may seem hard to accommodate within the traditional division of views within modern moral philosophy. On the one hand, though it carries echoes of Kantian notion of making law for oneself, it runs very much against the usual picture of Kant's emotionless notion of practical reason (see esp. Kant 1981 , 99; but cf. Sherman 1990). On the other hand, while the contrasting Humean picture accords emotions a necessary practical role, this seems to depend on undermining the role of practical reason. However, here as elsewhere an alternative model is provided by Aristotle (1984 : *Nicomachean Ethics*). Among other things, Aristotelian ethics suggests a picture of practical reasoning that includes not just the calculation of means to ends but also the determination of constituents of valued wholes. Aristotle had in mind happiness, of course, but one might fill in something shorter-term for particular emotions. Action expressive of an emotion has a certain value for the agent as the completion of a more integrated psychological whole: a fit between inner and outer states of oneself, as the original condition from which adult reactions developed.

For that matter, action on emotion counts as an instance of value-expressive behavior to the extent that emotions are felt evaluations: they amount to affective states assessable from the standpoint of the agent but also with evaluative thought content. So there are two layers of evaluation that we can appeal to here. My suggestion is not that either of them is final. Expressing one's feelings is certainly not always the most rational thing to do. For one thing, the evaluations that emotions register need not represent one's considered view of the situation. But there is always *pro tanto* reason for expressing emotion, in the way thinking that *p* is a *pro tanto* reason for saying “*p*,” even if it is ultimately defeated by countervailing reasons. That it manifests a state of mind is a consideration in its favor.

end p.217

The reasons that emotions provide according to this account are supposed to be justifying reasons, reasons in a normative sense—good reasons, at least as far as they go. There is some resistance to according emotions this role, in part because it seems to cross categories of emotion versus reason, but also because of its apparent implications for moral motivation. Motivation by moral emotion, if understood in the terms I have suggested, can seem suspiciously self-regarding—as if the agent's goal in acting morally were either to improve her state of comfort or just to express herself. However, if the content of emotion is a moral evaluation or an evaluation that has moral worth (such as one that recognizes the value of another person), I think this objection can be met, at least for any reasonably broad understanding of moral motivation.

One might want to say that emotions themselves can at best just heighten the force of nonemotional reasons, whatever reasons there are for the evaluative content of emotion, or for holding it in mind. But consider how undertaking an obligation by promising or the like typically augments such reasons as already may have been in force for acting in the way that the obligation prescribes. Affect adds a further factor for an agent to take into account, a psychological factor under limited control.

Within moral philosophy, attention to emotions as subjects of moral judgment has generally emphasized their role as elements of personal moral worth (see, e.g., Oakley 1992), on the assumption that their uncontrolled aspect makes them unfit subjects of moral requirement. Indeed, emotions are often invoked in explaining weakness of will. However, insofar as they involve more or less immediate comfort or discomfort, they can serve to modify the payoff structure of the situation prompting weakness of will by reversing temporal discounting (see Frank 1988 , Greenspan 1988 ; cf. Elster 1999). Their uncontrolled aspect can itself serve as a check against weakness of the sort that involves recalculating one's reasons when the time for action arrives. Emotion provides a reason for action that is itself somewhat recalcitrant to reason, though subject to indirect control via control over attention.

The idea of emotions as factors in practical reasoning might seem to set up a dilemma. On the one hand, if their rational role is analogous to that of belief, then they are reasons only in a subjective sense, not as facts about the world (including ourselves) but as purported representations of facts. My feeling in the consumer case that the store has treated me unfairly would seem to justify threatening them with a complaint only on the assumption that it reflects the truth of the matter. On the other hand, to the extent that we think of emotions as providing new facts that we need to consider in practical reasoning—my angry feelings in the consumer case as unpleasant feelings I ought to act to assuage—they seem to play an obvious but trivial role, in the manner of “brute” facts like the fact that one has a headache, viewed as a reason (say) for staying home from a disco.

end p.218

However, consider what it is like to hate the disco scene and to feel repugnance at the thought of being part of it. The evaluative and affective aspects of emotion (or the facts they represent or constitute) reinforce each other as part of a single reason for declining an invitation to a disco: it would be awful, hence I could not stand it, hence it would be all the more awful—and so on. This combined reason is not subject to either side of the

alleged dilemma, and it provides a firmer barrier to persuasion than either evaluation or affect alone could hope to do.

Philosophers who accept a desire/belief account of intentional action—a contemporary reading of Aristotle's practical syllogism that is often thought of as Humean—sometimes interpret the motivational role of emotions in terms of desire (see, e.g., Marks 1982 , Searle 1983 , pp. 31–33, Gordon 1987 , Wollheim 1991 , Green 1992). Emotions of course can give rise to desires—a desire to avoid discos, in the case just outlined. In the contemporary philosopher's sense of desires as simple "wants," which leaves out the element of emotional affect, there is a trivial sense in which this would have to be true of any emotions that result in action. However, emotions do not simply reduce to desires, or desire/belief pairs; we need to bring in affect to explain their full rational bearing on action. Nor need the reasons for emotions depend on the agent's (prior) desires: there may be reasons for feeling something—a moral requirement of compassion toward others or guilt at wrong action—that apply to us regardless of our desires.

It may concern some moral philosophers that the bearing of emotions on action according to this sort of account would depend on our actually feeling them. Whether we do or not is of course a contingent matter and not subject to complete control. However, presumably the general moral requirement that is satisfied by feeling a given emotion would also require (perhaps as a second-best alternative) that one muster whatever other motivational resources one can to perform the kind of act that typically results from that emotion: aiding those in distress, say, for compassion, or making up for past wrongs, in the case of guilt.

It is important not to confuse the moral or other practical reason to feel a given emotion—whether as a reason for action or just in itself, as part of a personal ideal of virtue—with the reasons why it is appropriate to feel. According to the perspectival account outlined here, the latter are analogous to evidence for rational belief, except that we do not assume it is irrational *not* to feel an appropriate emotion. At this stage, if rational warrant is all that is in question, we have something like a theoretical syllogism yielding sufficient but noncompelling reason for an emotion. Its conclusion serves, or can serve, as the first premise of a practical syllogism leading to action.

end p.219

7. Evaluative Affects

My use of the propositional attitudes approach to emotions has been directed in the first instance toward questions of rationality. This is usually taken as a later question, but I think that an answer to it can be read back onto the philosopher's traditional first question, about the nature of emotions, for an answer that does not pull quite so sharply away as judgmentalism did from the traditional approach to emotions in terms of feelings.

For recent attempts to reestablish the feeling theory, see Leighton 1985 , Robinson 1995 , and Pugmire 1998 . However, emotions can be feelings and still be propositional attitudes. The propositional attitudes approach need not be understood as assigning to emotions a separable element of propositional thought as distinct from feeling. My

suggestion is rather that emotional affect itself *renders* a judgment that can be stated in propositional terms—by us theorists, that is, rather than necessarily by the agent.
The view depends on assigning the affective element of emotion an intentional content. The assumption of intentionality at this level of basic feeling can sound mysterious, but in principle it is no more so than in familiar cases involving units of language and thought. Insofar as emotional comfort and discomfort does have a content, moreover, the affective element of emotion should not be equated with simple hedonic states of pleasure or pain. Instead of thinking of "content" on the metaphor of containment in a composite entity, the idea of emotional affect as pointing toward something outside itself (though internal to the composite entity, of affect plus evaluation, that we think of as the emotion) might be less misleading, with bodily gestures like pointing as the appropriate analogue.
The appropriate cognitive analogue, rather than linguistic meaning, would seem to be reference—as in the thesis from Brentano 1973 that made intentionality an issue, the "intentional reference of the mental." We need not assume that the relevant referential relationship is type-to-type—that a given sort of affect always refers to the same proposition—as cases of demonstrative or other indexical reference make clear.
For emotions like love and hatred whose surface structure is not propositional (one need not love the fact *that* the love object has such-and-such a feature), we can distinguish propositional components of emotion corresponding to its motivational role: discomfort that one is far from the love object, say, corresponding to the desire for closeness characteristic of love (Greenspan 1988). My own rather minimal assumption has been that the positive or negative aspect of a given feeling—understood in terms of motivational significance rather than hedonic tone—represents the positive or negative aspect of the corresponding evaluation. However, there are ways of further specifying this simple bipolar account, which was set up just for purposes of getting at the rational role of emotions.
end p.220

The evidence cited by Griffiths (1997) from Ekman 1992 —and ultimately from Darwin 1965 —groups our basic emotions into multiple "affect programs" that evolved initially to prepare the body for different modes of action: fight or flight, for instance, in the familiar cases of anger and fear. The element of feeling in these clusters of responses comes to be a sign of, and in that sense "about," the need for a certain kind of action. This natural referent of feeling would then be subject to social modification, including the sort of individual interaction with caretakers that de Sousa's account of paradigm scenarios takes as explaining the intentionality of full-fledged emotions.
This account might be extended to cover purely expressive action, without evolutionary function, possibly mediated by symbols based on social learning. A more mentalistic variant might include or substitute reference to an associated mental act such as directing attention—or simply the act of taking some considerations as reasons for action (cf. Scanlon 2002). At any rate, given the broad use of "judgment" in the emotions literature, the general suggestion that feeling itself can have an evaluative intentional content may be thought of as a friendly amendment to judgmentalism. There are parallels to it in the continental literature (cf. Brentano 1973 , 99, 223), though with emotions presented in

contrast with judgments. In general, the view that “affect evaluates” (as I would sum it up in slogan form) affords a way of accommodating empirical findings (as in Zajonc 1980, LeDoux 1996) that some cases of emotion, such as primitive fear responses, do not involve cognition (cf. Deigh 1994 , Robinson 1995). Cases of “gut feelings” recording good reasons for action suggest that even inarticulate emotions sometimes act in aid of rationality.

NOTE

For comments on earlier versions of this paper, I owe thanks to Erich Diese, Scott James, Stephen Leighton, and Jerrold Levinson.

chapter 12 THE RATIONALITY OF BEING GUIDED BY RULES

Edward F. McClennen

Intuitively, to be guided in one's choice of action by a rule is not simply to choose in conformity with that rule; it is also to choose to engage in or refrain from a course of action *for the reason that* it is required by, or prohibited by, a rule. In such cases the rule is thought to determine whether a given choice of action on our part is justifiable or acceptable. Of course, there is a substantial issue that needs to be resolved as to what it means to say that there exists a rule that functions in this way. In the case of the rules of a free, liberal, democratic society, however, it is customary to suppose that the rules that function in this way are those that one has accepted or to which one has made a commitment and perhaps even helped to formulate, directly or indirectly. In what is to follow I shall limit myself to a discussion of the issue of the rationality of being guided by rules to which one has made such a commitment. [1]

It would seem clear, of course, that not just any kind of rule could plausibly function in this manner and thus provide reasons for the choice of an action. As John Rawls (1955) argues in “Two Concepts of Rules,” what are sometimes called “maxims” or “rules of thumb” are not ordinarily thought to provide this sort of reason for action. Rather, they serve merely to summarize past findings concerning the application of some general choice-supporting consideration to a partic

end p.222

ular case. Such a rule, he argues, presupposes choice-supporting considerations and cases that are logically prior to, and thus can be described without reference to, the rule. Moreover, the person faced with choosing is always entitled to reconsider the correctness of what the maxim recommends or urges—by direct appeal to the underlying choice-supporting considerations. Such rules, then, appear to serve a purely heuristic function. But there are, on Rawls's account, rules that do provide reasons. One class consists of the rules of practices—social, commercial, legal, and political. Such rules are, he suggests, logically prior to the cases to which they are to apply, and those who participate in such

practices are not at liberty to decide for themselves on the propriety of following the rule in particular cases. Specifically, while the rule itself may be defended by appeal to various considerations—for example, various general, utilitarian considerations—those who are bound by the rule cannot take exception to it by direct appeal to those considerations. That is, one can explicitly invoke a "two-tiered" theory of justification for at least certain practices. One appeals directly to a rule to justify or defend the choice of an action, and while the rule in question is itself subject to a defense in terms of various supporting considerations, one cannot directly appeal to those considerations in order to justify departing from the rule.

Rules of this sort are, for example, involved in various sports and recreational games. By reviewing the history of the game of baseball, for example, one can, no doubt, offer some sort of explanation as to why the "three-strikes-and-you-are-out-rule" was incorporated into the game. Whatever those considerations were, you could not stand at the plate and start arguing with the umpire who has just called you out by that rule, urging that consideration of those underlying reasons supported your proposal that this time at bat you be allowed four strikes. Again, one may place a special, intrinsic value on accepting and being guided by certain rules—as in the case of dietary rules associated with a particular religion.

But what are we to say regarding the appeal to a two-tiered theory when it comes to rules the following of which is presumed to be of purely instrumental value, that is, productive of various kinds of benefits? The rationality of subsidiary rules within a utilitarian moral and/or political theory is a case in point. But, more generally, it is customary to defend various rules by appeal to what would benefit the individuals within a group who are trying to coordinate their actions in a manner that will prove mutually beneficial. In these kinds of cases we appear to face a fundamental dilemma.

Consider any rule *R* and suppose that *R* requires an agent *X* to do *A* in circumstances *C*. When *C* obtains, either there is a better option than *A* available or there isn't. If there were a better option, then it would seem to be irrational to perform the inferior option *A*. On the other hand, if there is no better option than doing *A*, that is, if doing *A* is supported by the balance of reasons, then *X* is justified in choosing *A*—not because *A* is required by *R*, but because rational

agents should always pick options that have no better rivals. In short, rule-guided behavior cannot be justified: either the rule gives the wrong result, in which case it is irrational to follow it, or it gives the right result, in which case guidance by the rule is irrelevant. [2]

This dilemma presents a challenge to the idea of rule-guided choice within the framework of a theory of rational deliberation in which personal, moral, social, or legal rules are regarded as having an instrumental value. It suggests, in effect, that it is odd for a rational agent to cite a rule as a conclusive reason for choice of an action and then acknowledge that the rule was heeded because of the benefits that rule-following brings. In the absence of attaching a special value to rule-following as such, how can the rule in this case play a role in justifying action? The dilemma is very robust, moreover, because it seems to turn

on nothing more than a very weak assumption, namely that it is irrational to choose an option that is judged inferior to another identifiable and available option.
To the dilemma as formulated there would appear to be three distinct responses. First, one can accept the conclusion that rules have no real standing in a theory of the justification of choices. That is, one treats the dilemma as a valid *reductio* and accepts both of the disjuncts above: rule-guidance is either irrational or redundant. This amounts to adopting an "error" theory of rules, according to which rational rule-guided choice of action is only apparent. [3] All in all, this seems to be an option of last resort, to which one would retreat only if one is convinced that all other possibilities have been exhausted. Second, one can adopt a *compatibilist* strategy and grant that rational agents should always perform the action supported by the best reasons but insist that rule-guidance can be very relevant to instrumental practical reasoning. The compatibilist assumes that every choice that the agent can make is associated with reasons for and against. These reasons can be compared and, most importantly, aggregated. The choice that is justified, then, is the one that is supported by the greatest balance of pro over con reasons. On this view, rules can provide pro reasons for acting in accordance with them by appeal to a variety of practical considerations, chief among which is that they can help us to coordinate our actions over time, that is, execute in a coherent fashion the plans we adopt, coordinate our actions with others, and save on deliberation costs. The compatibilist, then, denies that if the rule gives the right result, then the rule is irrelevant or redundant. Thus the compatibilist strategy allows one to preserve the idea that rational agents should always be faithful to the underlying reasons for action, while also providing a role for rules. The compatibilist is still committed to the view, however, that if allowing some rule to guide our choice was not supported in a given case by the balance of reasons, then it would be irrational to be guided by the rule. That is, the compatibilist holds the view that a particular instance of rule-guided choice is rational only if that choice is supported by the balance of reasons that apply to that case.
end p.224

The *third* response is to adopt a *revisionist* approach. The revisionist agrees with the compatibilist that rule-guided behavior does have a real place within an instrumental theory of justification, but one that forces us to revise the standard account of practical reasoning. The revisionist, then, denies that if the rule gives the wrong result it would be irrational to follow it. A typical revisionist strategy is to argue for some sort of hierarchical model of justification, in which a distinction is made between justifying the choice of rules and justifying the choice of an action within the framework of the rules that have been adopted. [4] That means, however, that on the revisionist account, the door is opened to the possibility of rule-guided rational action that is not supported by the balance of reasons applicable at the time the action is taken. Predictably, whenever a revisionist insists that it is rational to be guided by a rule, the compatibilist will argue that the reasons for following the rule actually tip the balance of reasons in favor of conformity to the rule. Compatibilists are inclined, then, to argue that revisionism is not a coherent position, in that it must sanction actions that are not supported by the balance of reasons.

But revisionism has attracted a significant following in recent years. One can trace this, in great part, to the influence of Joseph Raz, who argued that it might be rational to act in ways that are not recommended by the balance of reasons. On Raz's view, rules can constitute "second-order reasons"—*exclusionary* reasons—not to act on certain first-order reasons, where the latter are the sorts of reasons that lend themselves to the balancing approach. [5] When an exclusionary reason applies, the first-order reasons that are excluded by it are completely removed from the balance of reasons. Thus it would be possible that a balance of first-order reasons recommends an action but that the existence of an exclusionary reason requires the agent to disregard that recommendation and act in a different manner.

Raz insists that certain reasons for acting must be exclusionary in nature, for if they were just first-order reasons, to be considered along with other reasons, they could not perform the function we normally attribute to them. Consider the example here that one should in certain contexts economize on deliberation costs by employing a rule of nondeliberation. [6] If this were just a first-order reason not to deliberate, then the only way of knowing if this reason were outweighed in a given context would be to consider all of the first-order reasons for action—that is, to deliberate—but that would defeat the very purpose of the rule of nondeliberation. Again rules that secure coordination can be understood as providing exclusionary reasons, for if everyone acted on the balance of reasons as they each saw it, concerted action would be impossible.

Raz's work occasioned a lively exchange with a number of compatibilists. In part this constituted an in-house debate between Raz and a variety of other legal theorists. [7] But the dispute was more general. In part what the compatibilists argued was that any valid example of what Raz characterized as an exclusionary

reason could be treated simply as a first-order reason that was very weighty. It was also argued, however, that Raz had never really given an answer to the question of how one could be justified in not acting according to the balance of reasons. [8]

Quite independently of the revisionist line of inquiry that Raz initiated, the issue of the rationality of rule-guided behavior arose also among economists and decision theorists. In the second half of the twentieth century, the prevailing view among economists was that the concept of a preference ordering, which was central to microeconomics, must be given a behavioral interpretation in terms of how the agent chooses. To say that one prefers x to y is to say that when confronted with a choice between x and y one will choose x. On this way of thinking, it is plausible to suppose that if one chooses to be guided by a rule, then in so doing one simply reveals one's preference for being so guided. To be sure, such a preference seems odd within the context of a theory of consumer choice. More to the point, appealing to the idea that in choosing to be guided by a rule one is simply revealing one's preference for being rule-guided has something of the air of a purely ad hoc assumption. However, the theory of decision making based on the maximization of one's preference ordering over possible options was taken over by decision and game theorists (see Joyce, chap. 8, Dreier, chap. 9, Bicchieri, chap. 10, and Danielson, chap. 22, this volume), who were happily prepared to expand the concept of a mathematically representable preference ordering (a utility function) to any decision

problem that an agent might confront. Within this context, then, it might seem as if the concept of rule-guided behavior posed no theoretically interesting problem.

But in fact there were those who found themselves worried about this interpretive move. The most elegantly and powerfully formulated dissent came quite early on from the economist and philosopher Amartya K. Sen:

> A person is given one preference ordering, and as and when the need arises this is supposed to reflect his interests, represent his welfare, summarize his idea of what should be done, and describe his actual choices and behavior. Can one preference ordering do all these things? A person thus described may be "rational" in the limited sense of revealing no inconsistencies in his choice behavior, but if he has no use for these distinctions between quite different concepts, he must be a bit of a fool. (Sen 1977 , 336)

Sen's disaffection with this sort of account is especially due to what he regarded as its lack of a coherent account of the notion of choice behavior based on commitment—just the sort of behavior that appears to be involved when one is led, in virtue of accepting a rule, to regard the rule as providing a reason for choosing one way rather than another.

As it turns out, however, it was not just the facile assumption that commitment could be assimilated to a special kind of preference that engendered worries about how to understand rule-guided choice. In fact, the issue of the nature of rule-guided choice was forced upon economists by a special class of decision problems, in which one faced the far from anomalous problem that one's preferences now among various options that one would face in the future often turned out to be different from the preferences one would then have (that is, in the future) for those same options. That is, what brought the issue to a head was the problem of what came to be known as time-inconsistent preferences. The seminal work here is a 1955 article by R. H. Strotz, who argued that the characteristic way in which persons discounted the future lead to such time-inconsistent preferences. Strotz's own examples were primarily drawn from the area of personal decision making—as, for example, when one seeks to adopt a savings or dieting plan, only to find that, subsequently, one is tempted to depart from the plan. [9] But as Kydland and Prescott (1977) subsequently argued, the problem is also endemic to government policy making. [10]

The striking thing about the ensuing analysis was that the standard theory of rational deliberation turned out to have only limited resources with which to deal with the issue raised. If one now wants that in the future one adheres to a rule, but one worries that the future self will want instead to disregard the rule, then on the standard way of thinking one has only one of two strategies available. One can *precommit*, that is, tie one's hands now so that one will be powerless to deviate from the rule in the future (a strategy that is a close cousin to that of agreeing to an enforcement mechanism that will compel one to behave a certain way in the future, upon pain of suffering an unacceptable outcome), or if one lacks the means to control one's future self in this manner, one can recognize now that planning to adhere to the rule in the future is futile, for it would be to adopt a plan that one realizes will never be executed. Such strategies, however, could be characterized as second-best strategies, for they require the self to forego adopting plans that would be beneficial, or to expend scarce resources to ensure that their future selves are bound to do their part.

The problem of time-inconsistent preferences is closely related to another problem with which economists and game theorists have had to grapple, again with somewhat limited

resources, namely the free-rider problem. It is characteristic of many mutually beneficial arrangements between persons that what each participant stands to gain from is that others conform to the rules defining the arrangement, and that their own conformity to those rules is a cost to them. In such cases, each will be disposed to encourage others to conform but be disposed to avoid conforming whenever possible. The problem can arise, of course, even in the very simplest of contexts, such as contractual or reciprocal arrangements in which goods are exchanged, but there is a temporal separation between the performances expected of each. Both parties may be interested in the exchange, but the person who must perform first has to contend with the problem that if he or she performs, the other party may now have no motive to reciprocate. Once again, the standard theory of rationality has limited resources with which to deal
end p.227

with this problem. As in the case of the self that has to deal with the rational dispositions of its own future self, those who have to deal with the rational dispositions of other selves are limited to either forgoing certain arrangements that would be mutually beneficial or expending scarce resources to ensure that others do their part—that is, by providing sanctions against free-riders.

It is in this context that revisionist views could be said to have come into their own. One can start by noting first that Raz's exclusionary reasons account is applicable to cases of this sort. To commit to a personal rule binding one's own future self can be thought of as creating a second-order reason not to act on the first-order temptation one will subsequently come to have to depart from the rule. If such "temptations" were not excluded from consideration, the balance of first-order reasons—one's preference in the balance—might well be to depart from the rule; but given the exclusion of the "temptation" the balance of reasons may well favor adhering to the rule. Similarly, to the extent that a number of persons were able to approach cooperative situations from an exclusionary reasons perspective, they would be able to execute useful arrangements without having to employ expensive enforcement devices.

Raz's exclusionary reasons account does not exhaust the resources of a revisionist view, however. There are, in fact, two other revisionist proposals that have more recently been made. One alternative involves appeal to what could be characterized as a *constraint model*, in which one thinks of commitment to a rule as a volitional act that alters the feasible set of future alternatives. Isaac Levi and Scott Shapiro have defended versions of this approach. [11] On this view, making a commitment to a rule involves an alteration in what one regards subsequently as the set of feasible alternatives from which choice is to be made. More specifically, making a commitment is to be understood as causally leading one to no longer regard certain actions as feasible. On this account, an agent is being (instrumentally) guided by a rule *R* if and only if (1) the agent conforms to *R*, (2) recognition that *R* applies constrains the agent to conform to *R*, making nonconformity infeasible, and (3) absent the constraint, the agent believes that he or she might not have conformed to *R*.

The constraint model of rule-guided behavior has the great advantage that it offers an unproblematic account of the rationality of being guided by a rule, since rationality has always been taken as matter of how choice is to be made from among a *feasible* set of

alternatives. But the problem for this approach is how to explain what has to be regarded as a special sort of feasibility—how to account for the rational agent coming to regard, by an act of will, certain logically feasible (and attractive!) alternatives as no longer practically feasible.

A third approach involves appeal to a *resolute choice* model. Both I and David Gauthier have explored versions of this approach. [12] On the resolute model, one is to suppose that—in contradistinction to the standard way of thinking, where present choice can only be adjusted to what one expects to choose in the future—a

end p.228

rational agent can choose to constrain future choices to what one has decided in advance, to a prior commitment to a rule. On the version of this approach that I originally put forward, commitment is to be understood as a volitional act that (causally) transforms one's subsequent preferences, that is, one's future preferences regarding certain future actions. [13] This version of the resolute model, then, could be said to share with the compatibilist position the view that a particular instance of rule-guided choice is rational only if the balance of reasons applying in that particular case supports that choice. If the resolute model manages thereby to avoid the constraint model's problem of having to invoke a special notion of feasibility, it nonetheless faces a parallel problem: explaining how the commitment to a rule at one point in time can transform one's future preferences.

It should be noted that there are a number of things that all three of these revisionist models have in common. First, they each view agents as capable of volitional acts that have *temporal thickness*, with the present self, at least in certain circumstances, being able to exercise rational authority over the behavior of the future self and thereby justifying future obedience to the rule adopted. Second, each of these ways of thinking about rule-guided choice preserves our sense that in such behavior there is a tension between commitments made and preferences at the moment of choice. On Raz's account, the temptation to break the rule, while excluded from deliberation, is still present. On the constraint model, there is still a tension between the commitment made and present preference, for the temptation to break the rule has once again not been removed; it is just that it is no longer viewed as a feasible option. On the resolute choice model, even though one's overall preference, at the subsequent point of choice, is for conforming to the rule, this does not mean that one is not tempted to break the rule—for were other things equal (which they are not, since one has made a commitment) one would prefer to break the rule, and the awareness of this fact preserves the sense of a tension between commitment and choice. Third, in each case, as the above examples make clear, the analysis need not be confined to cases of personal rules. There is, on each of the three accounts, a natural extension from the case of the agent faced with making decisions over time to the case of different agents trying to coordinate their actions. [14]

From an analytic point of view, one important issue is obviously whether one or another of these revisionist models can be shown to be superior to the others, in respect to giving a satisfactory account of rule-guided choice. And there is also an unresolved question whether these three models exhaust the possibilities. The work of Ainslie on personal rules seems to combine elements of both the constraint and the resolute models, but this may not do justice to what is a very original and interesting account of personal rules. [15]

Ainslie's work also brings to the subject extensive empirical findings drawn from the fields of physiology and psychology, and it is clear that future analysis could greatly benefit from more integration of such empirical work with the philosophical, legal, and decision-
end p.229

theoretic perspectives that I have been discussing. So also there is much to be gained by examining the recent work of institutional theorists, work that has greatly enriched our understanding of the role of rules in social, political, and economic activities. [16]

I want to put the issue of the best way to characterize rule-guided choice to one side, however, for I think there is an even more pressing problem. This is a problem that arises in virtue of an objection that can be raised to any of the models discussed by adapting an argument of Michael Bratman's regarding the role that plans or intentions play in our deliberations. Intuitively, we want to say that intending to do something provides a reason for action. But one has to be careful here. Bratman (1987 , sec. 3.3) argues that, as stated, the view is too strong. Even though the idea of an intention-based reason allows us to take seriously the way in which one sees one's prior intentions as directly relevant to the conclusions of one's further practical reasoning, it appears to sanction an unacceptable form of bootstrapping. Suppose that at some subsequent point in time, one comes to *reconsider* one's previous intentions. It would be an unacceptable form of bootstrapping, Bratman argues, to suppose that the mere fact that one previously intended (as distinct from one's evaluation now of the consequences of acting on that intention) carried any weight. Adapting this point to the present discussion of a commitment to rules, the issue is how to understand the role such a commitment plays in deliberation without falling into the unacceptable position of letting the mere fact of having made the commitment create a distinct reason in favor of guiding one's choices by reference to the rule.

Bratman thinks that the way to avoid such bootstrapping is to treat the adoption of a plan not as creating new preferences, but as merely constraining the range of actions available for subsequent choice. And Gauthier (1996 , 218), in commenting upon Bratman, thinks that this is the way to deal with the bootstrapping problem. This would seem to suggest that each is implicitly inclined to think that a constraint model is the most appropriate revisionist model. It can be noted, moreover, that in Gauthier's own explorations of "resolute choice" he clearly dissents from the suggestion, in my book, that the best way to understand resolute choice is by appeal to the idea of a commitment that causes one to have different preferences for the options available at some subsequent choice point—to the idea of what I characterized there as "endogenous preference changes." [17]

My own sense, however, is that the issue goes deeper. It is not just a problem that arises within a resolute model of commitment, that is, one that invokes the idea of endogenous changes in preference. The issue is really one of intelligible reasons, and Raz's exclusionary reasons model, as well as Levi and Shapiro's constraint model, strike me as posing the same sort of problem. That is, the more general question is just how the agent's commitment to a rule could provide reasons for the agent to be guided by that rule. [18] The problem is that it seems odd to suppose, within an instrumental framework, not only that a commitment
end p.230

to a rule could create a new situation in which the balance of preferences would be tipped in favor of adhering to the rule, but also that it could yield a revised view of what is feasible or generate some hierarchical structure of first- and second-order reasons, with the latter constituting exclusionary considerations. Bratman's bootstrapping worry is, I think, at root a worry about how to provide a role for rules within an instrumental theory of rationality without finding oneself faced with a "bootstrapping" move that leaves one advocating "rule worship" or "rule fetishism." [19] Why should we suppose that telling the story in terms of constraints on the feasible set, or in terms of second-order reasons, rather than in terms of (new) preferences over those same options, somehow solves the problem of bootstrapping?

The key to understanding how to avoid the bootstrapping charge lies, I suggest, not in how we characterize commitment to a rule, but in what story we can tell about what *drives* or *motivates* the commitment. And here it will prove useful to return and explore one paradigm case of rule-guided choice—namely situations in which the utilization of a rule of nondeliberation (i.e., a rule of nondeliberative nonreconsideration) serves to reduce decision-making costs. [20] I noted above that Raz's exclusionary reasons approach can be illustrated by this kind of case. The argument was that if this were just a first-order reason not to deliberate, then the only way of knowing if this reason were outweighed in a given context would be to consider all of the first-order reasons for action, that is, to deliberate—but that would defeat the very purpose of the rule of nondeliberation. The argument, then, is that there is a cost-savings consideration that drives the commitment, savings that cannot be achieved except by not being disposed to reconsider continually the pros and cons of applying the rule in each subsequent case that arises.

Having such a rule of nonreconsideration in place for certain classes of cases will, then, clearly save on costs of deliberating. Of course, as Bratman makes clear, any such rule will be subject to all sorts of exceptions. That is, the savings in deliberating costs can be outweighed by the severity of the negative consequences that can flow from not deliberatively reconsidering in certain contexts. Presumably, then, the commitment to nondeliberative nonreconsideration will not be unconditional: certain conditions will serve to trigger deliberative reconsideration.

But invoking Raz's notion of economizing on decision-making costs by having a second order exclusionary rule is not the only way one can rationalize a rule of nondeliberate nonreconsideration. We could also appeal to a constraint model and characterize being guided by such a rule as a case in which the option of reconsideration, which is typically available to us at subsequent points in time, is simply (for at least certain classes of cases) treated as a nonfeasible option. That is, we could interpret commitment to such a rule in terms of the constraint model. Similarly, one could interpret commitment to such a rule as a case of acting in a resolute fashion—that is, deciding, for reasons again of cost savings, to adopt a rule of nonreconsideration, which would lead one to prefer to not deliberatively reconsider some decision reached previously, unless, say, the real possibility of significant negative consequences was thrust upon one.

Whichever story is told, it is clear that each of these ways of describing the rationality of a rule of nondeliberative nonreconsideration commits one to a revisionist program. That is, to accept such a rule and be guided by it is to put oneself in a position in which the

balance of reasons could turn out, if only one had deliberatively explored them, to favor reconsideration. That is, the rule counsels one to choose in a manner that will not always ensure that one chooses in accordance with the balance of reasons that arise within the context of a particular act of choice. Thus accepting such a rule cannot be rationalized within the framework of a compatibilist position.

Moreover, whichever one of the three ways of interpreting a revisionist account one employs, in each case it would appear possible to avoid the bootstrapping charge. This is because the commitment to the rule, on any of these accounts, is based on a pragmatic consideration—on a consequentialist concern to hold down costs. [21] What drives the argument, then, is not the mere fact of making a commitment to the rule of nonreconsideration, but the cost-saving consideration behind the making of that commitment.

Now, against the background of these considerations, consider once again the way in which the standard theory of rationality deals with the self over time (and with persons who have to interact sequentially with one another). [22] First, the standard theory is really distinguished by a particular view of what it means for the self to deal with its own future selves, and, correspondingly, for the person to deal with other persons. This view, I want to argue, is essentially an *autistic* view of how the choice behavior of other selves or other persons is to be worked into the analysis. On this way of thinking, the choices made by one's own future self, or other persons, constitute so many conditioning factors—variables—whose value a rational agent must try to estimate, not for the sake of coordinating one's own choices with those other choices, but simply with a view to enabling the present self vis-à-vis future states of the self, and the self vis-à-vis other persons, to maximize its own utility function. That is, the perspective that is presupposed is that these conditioning factors are to be treated in the same way that other factors—states of nature—are to be treated: as variables whose value must be properly estimated if one is to achieve effectively maximum preference satisfaction. These are models, then, in which the self ends up taking the choice behavior of the future self (or other selves) as *given*, as something to which one's present self must adjust.

In the case of sequential choice, however, the standard way of conceiving the constraints on a prior decision maker are a bit more complicated. This is because the later self (or the other person who chooses subsequently) can be expected to maximize from its own perspective, against the known choice of the earlier self or person. The problem for the earlier self, then, is that it must contend with a

subsequent self that will not have to make a choice until the earlier self has chosen, and will then be in a position to maximize from its own perspective. Thus, the person who, in the morning, resolves to adhere strictly to a diet has to face the prospect that he or she will, in the evening, have a preference for taking a helping of some forbidden dessert. Correspondingly, a parallel problem is posed for two farmers, *A* and *B*, where *A*'s crops will be ready for harvest earlier than *B*'s. [23] *B* would be willing to help *A* first, if only, subsequently, *A* would reciprocate. But *B* has to contend with the real possibility that *A* will, when the time comes, refuse to reciprocate. Given this, the task of the earlier self (or first person to offer to help) is to choose so that the final outcome, given that the

subsequent self (or other person) maximizes against what the earlier self (or person) has done, is maximally preferred by the earlier self (or the first to offer to help). Thus, the standard model locks us into a view of mutual maximizing adjustments of each to the other, with the additional stipulation that given the existence of a natural temporal order, the process of mutual adjustment is structured in a certain way: the later self (or person) merely reacts to the earlier self (or person); the earlier self (or person) reacts to the predicted reaction of the subsequent self (or person). [24]

Now it is a well-established fact that such a form of strategic interaction between earlier and later selves, or between two persons who choose sequentially, will typically generate an outcome that is less advantageous for each of the selves involved than one that would be possible to achieve if the selves involved were to coordinate their choices cooperatively. In the language of the economist, the outcome will be Pareto-suboptimal. [25] To return to our examples, the would-be dieter must employ some device to "tie the hands" of his subsequent self—a device whose employment means that his goal is achieved in a more costly manner (than if he could simply resolve to keep to his diet). Correspondingly, the two farmers will either fail to help each other, to the detriment of both, or they will have to expend additional resources to ensure in some fashion or other conditions of trust. The farmer's problem, then, is simply a sequential version of the standard Prisoners' Dilemma game, in which one player chooses and then the other player chooses in full knowledge of how the first has chosen. [26] One can escape from the resultant suboptimality of the outcome only by replacing the kind of strategic interaction just described with something radically different, by each self approaching the sequential decision problem as a coordination problem. [27] That is, effective coordination—coordination that promotes the interests of both—will require that *each* restrain her choices and not straightforwardly maximize.

It would be misleading, then, to suggest that the revisionist models turn upon the idea that commitment to a rule is just a regimentation of subsequent choice to an earlier choice of a rule. What we really seek is a model in which regimentation in either direction is replaced with some sort of cooperative form of interaction. More specifically, what the resolute model (or any other viable revisionist

model) must make clear is that rational interaction can and should proceed on other grounds than each maximizing against what the other self or person has chosen or can be expected to choose. That is, what could rationalize an alternative form of interaction would be precisely the possibility of the mutual gain to be achieved thereby.

But in this very consideration of mutual advantage, I suggest, one can find just the kind of practical perspective that could motivate the parties—the sequence of selves or the different persons—to make a commitment to act in a rule-guided manner. [28] If we suppose that rational selves or persons will be disposed to coordinate when so doing works to their mutual advantage, we have an account of how such selves or persons could make a commitment to a rule. And this will be a commitment that is fully consistent with a practical or instrumental perspective. The later self (or the person who chooses subsequently) can be understood to choose in a rule-guided fashion because this is a strategy from which it, no less than the prior self (the person who chooses first), stands to

benefit—even though, of course, as the argument goes from the perspective of the standard theory, the subsequent self (or person) has a reason to depart from the rule at that point. The point is that even though the subsequent self (or person) could do even better by departing from the rule, the level of gain to be achieved by nonetheless adhering to the rule is greater than what the subsequent self (or person) could have hoped to have achieved in the absence of its willingness to be guided by the rule. [29] Its reason for being guided by a rule in the choice it makes is, then, a thoroughly practical, instrumental reason: absent such rules, which serve to coordinate choice, *each* party to the interaction stands to do less well. [30] What is invoked in such a way of thinking is not some bootstrapping move in which commitment to a rule makes us liable to objectionable rule-fetishism, but rather simply a more global or holistic way of thinking about consequential benefits. [31]

To see this, compare the following two possible worlds. In the first, each of the members of some society of persons is disposed to cooperate with one another as a matter of principle—that is, they do so without having to employ any system of mutual surveillance and enforcement. For the sake of the argument, we may simply suppose that in the society in question there is a cultural tradition to this effect. Now imagine a second possible world just like the first, but where such cooperation is possible only so far as resources are expended on surveillance and enforcement devices that serve to provide persons with sufficient incentive not to "free ride" on the cooperative efforts of others, and thus to stabilize the cooperative arrangement. On the assumption that the costs of such devices are shared among all, it is plausible to suppose that everyone is better off in the former than in the latter world. In each, cooperation is ensured, but in the latter only by the expenditure of resources that do not have to be expended in the former. To be sure, in the former world one might imagine that some come to realize that if they put to one side their cultural habits, they could do even better. But that

world is not stable. It will be rational for those who are thus taken advantage of to agree to expend resources to bring the free-riders into line. Of course, so long as most continue to cooperate as a matter of principle, the surveillance and enforcement costs will be less than they would be in a world in which all require such inducements. But so long as the inducement system is reasonably effective, and costs are shared, all would still do better to participate in the first possible world described above than in this modified version of that world.

What remains, now, is simply to ask, given the consequential superiority of the first possible world to the alternatives in question, whether rational beings might not deliberatively adopt such an arrangement. On the standard theory the answer is "No!" But does it really make sense to suppose that the benefits that each stands to realize by living in the first possible world can be secured only when, as a matter of pure cultural history, such cooperative norms are already in place—that it cannot be something that rational beings could bring about by any sort of deliberative process? That the standard theory must deny this possibility seems to me a telling argument against it—and one that proceeds by citing what are clearly consequential considerations. [31]

As this last remark is designed to suggest, if we now return once again to the issue of bootstrapping, it should be clear that this sort of way of understanding rule-guided choice is not subject to Bratman's criticism. However we choose to interpret being guided by a rule—whichever of the three revisionist models we employ—the manner in which being rule-guided is rationalized makes it clear that there is no bootstrapping going on. It is interesting to note, moreover, that the case we previously discussed, in which one is guided by a rule of nondeliberative nonreconsideration, can be fully rationalized in terms of this expanded and more holistic way of thinking about consequences. Indeed, were the costs to be saved by the adoption of such a rule credited only to the previous self, one could complain that this provides an inadequate pragmatic rationale for the adoption of such a rule. But the interesting thing about cost savings that the self can realize over time is that they are savings that can be carried forward and thus benefit the later self as well as the earlier self. Similarly, the cost savings that can be achieved by persons being able to coordinate their choices without having to employ expensive surveillance and enforcement devices are savings that not only can typically be shared among the participants, but plausibly must be shared, if what is clearly a much more rational form of interaction is to be realized.

NOTES

In preparing this article I have adapted a certain amount of terminology and framing of points from an encyclopedia entry that Scott Shapiro and I jointly authored (Mc
end p.235

Clennen and Shapiro 1998). When utilizing material from a joint undertaking, one is bound to be especially concerned to acknowledge when the borrowed material is primarily the work of the other author. Let me just say, in respect to the material I have borrowed, that the elegant precision with which many of the points and phrasings were made convinces me that they must have been due to Scott, and I am grateful, then, for his having given me permission to draw from what is officially our joint work.

1. The argument to be explored here could easily be adapted to address the larger issue of the place of commitment in a theory of rationality, as posed, for example, in a seminal article by Amartya K. Sen (1977). I shall content myself with developing the argument as it applies to the more special case of making a commitment to be guided by a rule.
2. A version of this argument is sometimes invoked to show that rule-utilitarianism must collapse into act-utilitarianism. Conversely, the line of argument I shall be exploring in this paper implies that that rule-utilitarianism is, at least with respect to this consideration, a perfectly coherent theory. I shall not explore this implication here, however, primarily because I think there are other substantial objections that can be raised against rule-utilitarianism.
3. Goodwin was perhaps the first to embrace this conclusion, in Goodwin (1793) 1971 . More recently, Wolff opted for a similar position in Wolff 1968 .
4. See Rawls 1955 , Wasserstrom 1961 , and Hooker 2000 .

5. This theme finds expression in a number of Raz's works, including Raz 1979 , 1986 , and 1990 .
6. Cases of this sort are also analyzed with much care and insight in Bratman 1987 .
7. See, for example, Green 1988 , Moore 1989 , Alexander 1990 , and Hurd 1991 . Raz offers a response to some of his critics in Raz 1989 .
8. See Moore 1989 and Schauer 1991 .
9. This kind of case was also the focus of a series of articles by Thomas Schelling. See Schelling 1984a , 1984b , and 1985 . It is also the focus of Elster 1984 and Ainslie 1988 , 1992 , and 2001 .
10. See also Yaari 1977 for a very comprehensive analysis of the logical structure of such problems.
11. See in particular, Levi 1986 . Scott Shapiro, who studied with Levi, developed this idea at great length in his dissertation. See Shapiro 1996 ; see also McClennen and Shapiro 1998 .
12. The term “resolute” was first used in this way in McClennen 1985 . I subsequently returned to this topic in a number of other articles, including McClennen 1988a and 1990 . McClennen 1988b is devoted to exploring the relation between this idea and Gauthier's idea of constrained maximization, as developed in Gauthier 1986 . In McClennen 1993 I take up the question of the relation between the idea of being resolute and the conception of planning developed in Bratman 1987 and 1992 . In McClennen 1997 I try to develop a less formal conception of resolute choice than the one contained in my book. During the late 1980s and 1990s, however, Gauthier himself contributed significantly to the analysis of being resolute in a series of articles. See, in particular, Gauthier 1988–89 , 1990 , 1994 , 1996 , 1997a , 1997b , 1998a , and 1998b . What is especially important about Gauthier's own explorations is that he has, I think, made it clear that the resolute model may not apply to games involving threats rather than assurances, and that even in the case of assurance games, there are limits to the model that must be acknowledged when chance events are also involved. My sense is that these limitations may also have to be acknowledged in the case of the other two revisionist models that I have been discussing.
13. See McClennen 1990 , sec. 12.7.
14. Indeed, the possibility of moving back and forth between the case of adopting a personal rule that is to be regulative of ones future choices and cases in which the rule in question serves to coordinate the choices of distinct individuals is implicit in the economist's very conceptualization of deliberation. To suppose that the earlier and the later self have different preferences—and hence different utility functions—is to suppose that the deliberative problem is essentially that between two different individuals who must interact in some fashion or other with one another.
15. For a brief introduction to Ainslie's account, see Ainslie 1988 . Ainslie 1992 and 2001 contain much more detailed accounts.
16. See here, in particular, Brennan and Buchanan 1987 and North 1990 .
17. McClennen 1990 , 214–15.
18. Recall that the issue here is how to understand such revisionist views within the context of an *instrumental* account of reasoning. That is, the issue is not whether it is possible to enjoy simply acting in conformity with a rule, or to attach intrinsic value to so acting.
19. This issue is explicitly raised in Bratman 1992 .

20. The terminology here is from Bratman 1987 , chap. 5.
21. I am using “consequentialist” in its most catholic sense here (not in the very narrow sense in which many twentieth moral philosophers have used it), to describe the position that all the consequences that flow from an action, for all the persons involved, are to be brought, and brought equally, into consideration—thereby invoking some version or other of a utilitarian rule of the maximization of the sum of benefits. On the other hand, I do suppose that if, on other grounds, some version of classical utilitarianism were defensible, the argument of this paper could be brought to bear in support of the idea of a *rule* version of that utilitarianism.
22. In what follows I will concentrate on the case in which cooperation with other persons involves a series of sequential acts, rather than the case where all choose simultaneously. But the argument to be developed can be extended to the case of simultaneous choice as well.
23. See Gauthier 1994 , 692.
24. In the literature on game theory this gives rise to the requirement that the sequence of choices made by rational persons (or rational time-defined moments of an individual person) satisfy what is known as the subgame-perfect equilibrium condition. For an elementary discussion, see Kreps 1990 , chap. 4. In McClennen 1990 , sec. 15.1, I argue that this condition is forced upon us by a *separability* assumption that mandates viewing the decision problem at any point in a decision tree as if it were a new decision problem involving only what is still open to agents to choose from that point onward. It is just this condition that gives rise to the idea that sequential choice problems can be analyzed by backward induction—by solving the last stages of a sequential problem first and then moving progressively backward through the tree until one comes to the first point at which choice is to be made. My own account of resolute choice was designed to break what I regarded as the stranglehold that this separability assumption and the related notion of backward induction had on the analysis of sequential choice problems—

a stranglehold that generates the profoundly implausible conclusion, for example, that under conditions of common knowledge, when two persons play a Prisoners' Dilemma game a known finite number of times, the only rational solution is for each to confess each and every time! The role that the separability assumption plays is discussed in length in McClennen 1997 . An especially lucid and accessible discussion of this assumption is also to be found in Gauthier 1996 .
25. Suboptimality in the case of the self over time means that *each and every* time-defined self is less well off than it would have been under some alternative arrangement. Both Kydland and Prescott 1977 and Yaari 1977 emphasize the suboptimality of the arrangements that can be realized within the framework of the standard theory of rationality.
26. If one considers the problem of two or more persons choosing simultaneously rather than sequentially, a similar result obtains. In terms of the standard theory of rational deliberation, the subgame perfect equilibrium condition is now replaced by the Nash equilibrium condition, and players who are fully rational (in the standard sense) will end up at an equilibrium outcome that is suboptimal, as the traditional version of Prisoners'

Dilemma makes clear, whether they simply do not trust each other and choose accordingly, or they expend resources to establish conditions of trust.
27. On the other hand, only in very special cases will such "games" constitute pure coordination games in which there is no conflict of interest. When the "game" is of this type, of course, it will have a trivial solution in virtue of its sequential structure. What makes for a pure coordination problem is that there are multiple solutions that are equally good, and the parties are unsure of which one to coordinate upon. But given a sequential structure, the party that goes first can thereby provide an appropriate signal as to which point to coordinate upon.
28. The distinction here would seem to mirror the distinction, to be found in Atiyah 1989 , between taking a contract as binding simply in virtue of one's having given one's word (simply made a commitment) and in virtue of the benefit that one expects to derive thereby. In opting for the latter conception, I am supposing that one's commitment itself derives from some consideration and does not just arise as a result of a pure act of will. And that also means that the consideration is relevant to the issue of whether the subsequent self is bound to the arrangement.
29. Because one still foregoes the additional gain that could be realized by deviating from the rule, the balance of reasons could still be said to support such a deviation; to appeal, then, to this rationale for being guided by rules is to appeal to a revisionist account.
30. This, one can insist, is a very different consequentialist consideration than, say, the consideration that one risks punishment or other losses if one is discovered to have not adhered to the rule. Again, the reader needs to bear in mind that I am using "consequential" in a broad sense: the argument does not turn on invoking the utilitarian ideal that the sum of benefits to all those who participate is greater.
31. Elster (1984) briefly alludes to just such a more holistic or global way of thinking about benefits, as it would apply to the problem of the self over time. The self that does not always, at each point in time, take the direction that leads upward will be able to avoid the problem of local gradient maxima in his or her path that are inferior to the gradient maxima to be found further on. That is a consequentialist consideration or reason in support of being, in effect, resolute. However, Elster retreats from the ex
end p.238

pected conclusion that rational agents should adopt such an approach, arguing instead that such self-restraint is not something that (imperfect?) human deliberators can achieve. But he fails to explain just why this lies beyond the power of mortals. Thus he argues that Ulysses is less than perfectly rational, because he has to resort to the device of having himself tied to the mast. Still, he goes on to insist, "he is capable of achieving by indirect means the same end as a rational person could have realized in a direct manner" (Elster 1984 , 36). The implication remains, it would seem, that he is not fully rational because he can achieve his end (avoiding the Sirens and continuing on home) only in a more costly manner. But, then, one must insist: What holds Ulysses back from the more efficient approach?
end p.239

chapter 13 MOTIVATED IRRATIONALITY

Alfred R. Mele

The literature on motivated irrationality has two primary foci: action and belief. In the former sphere, akratic action—action exhibiting so-called weakness of will or deficient self-control—has received pride of place. Philosophical work on motivated irrational belief includes, but is not limited to, work on self-deception. The primary topics of this chapter are akratic action and motivationally biased belief.

1. Akrasia, Self-Control, and Strict Akratic Action

The classical Greek term *akrasia*is formed from the alpha privative and *kratos*—strength or power. The pertinent power is the power to control oneself. Hence, *akrasia*is deficient self-control. Self-control, in this sense, is, very roughly, a robust capacity to see to it that one acts as one judges best in the face of actual or anticipated competing motivation. [1] The trait may be either regional or global (Rorty 1980a). A scholar who exhibits remarkable self-control in adhering to the

end p.240

demanding work schedule that she deems best for herself may be akratic about smoking. She is self-controlled in one region of her life and akratic in another. Self-control also comes in degrees: some self-controlled individuals are more self-controlled than others. People with global self-control—self-control in all regions of their lives—would be particularly remarkable, if, in every region, their self-control considerably exceeded that of most people.

In Plato's *Protagoras,*Socrates says that the common view about akratic action is that "many people who know what it is best to do are not willing to do it, though it is in their power, but do something else" (Plato 1953 , 352d). Here he raises (among other issues) the central question in subsequent philosophical discussion of *akrasia:*Is it possible to perform uncompelled intentional actions that, as one recognizes, are contrary to what one judges best, the judgment being made from the perspective of one's values, principles, desires, and beliefs? More briefly, is *strict*akratic action possible (Mele 1987 , 7)? Relevant judgments include judgments in which "best" is relativized to options envisioned by the agent at the time. In strict akratic action, an agent need not judge that a course of action *A*is the best of all possible courses of action open to her then. She may judge that *A*is better than the alternatives she has envisioned—that it is the best of the envisioned options. If, nevertheless, in the absence of compulsion, she does not *A*and intentionally pursues one of the envisioned alternatives, she acts akratically. It is a truism that a perfectly self-controlled agent would never act akratically. So akratic action, if it is possible, exhibits at least imperfect self-control. [2]

The judgment against which an agent acts in strict akratic action is what I call a *decisive*judgment. An agent's judgment that *A*is the best of his envisioned options at a

given time is a decisive judgment, in my sense, if and only if it *settles* in the agent's mind the question of which member of the set is best (from the perspective of his own desires, beliefs, etc.)—and best not just in some respect or other (e.g., financially), but without qualification. [3] Ann judges that *A*-ing would be morally (or aesthetically, or economically) better than *B*-ing and yet, in the absence of compulsion, she intentionally *B*-s rather than *A*-s. In *B*-ing, Ann need not be acting akratically, for she may also judge, for example, that *all things considered, B*-ing would be better than *A*-ing.

A feature of paradigmatic strict akratic actions that typically is taken for granted and rarely made explicit is that the judgments with which they conflict are *rationally* formed. In virtue of their clashing with the agent's rationally formed decisive judgment, such actions are subjectively irrational (to some degree, if not without qualification). There is a failure of coherence in the agent of a kind directly relevant to assessments of the agent's rationality. [4]

The occurrence of strict akratic actions seems to be an unfortunate fact of life. Unlike many such (apparent) facts, however, this one has attracted considerable philosophical attention for nearly two and a half millennia. A major source of the interest is obvious: strict akratic action raises difficult questions about connection between evaluative judgment and action, a connection of paramount importance for any theory of the explanation of intentional behavior that accords evaluative judgments an explanatory role.

Matters are complicated by our having—both in various theoretical approaches to understanding action and in ordinary thought—a pair of perspectives on the explanation of intentional action, a *motivational* and an *intellectual* one (Mele 1995 , 16–19; Pettit and Smith 1993). Central to the motivational perspective is the idea that what agents do when they act intentionally depends on where their strongest motivation lies then. [5] This perspective is taken on *all* intentional action, independently of the species to which the agents belong. If cats, dogs, and human beings act intentionally, the motivational perspective has all three species in its sights. The intellectual perspective applies only to intellectual beings. Identifying minimally sufficient conditions for membership in the class of intellectual beings is an ambitious task, but it is clear that the work of practical intellect, as it is normally conceived, includes weighing options and making judgments about what it is best, better, or "good enough" to do. Central to the intellectual perspective is the idea that such judgments play a significant role in explaining some intentional actions.

Many philosophers seek to combine these two perspectives into one in the domain of intentional human action. One tack is to insist that, in intellectual beings, motivational strength and evaluative judgment always are mutually aligned. Socrates seemingly advances this view in connection with his thesis that people never knowingly do wrong (Plato 1953 , 352b–358d). Theorists who take this tack have various options. For example, they can hold that judgment causally determines motivational strength, that motivational strength causally determines judgment, or that judgment and motivational strength have a common cause. They can also try to get by without causation, seeking purely conceptual grounds for the alignment thesis.

The apparent occurrence of strict akratic actions is a problem for this general tack. The motivational perspective is well suited to akratic action: when acting akratically, one presumably does what one is most strongly motivated to do then. But the intellectual

perspective is threatened; more precisely, certain interpretations of, or theses about, that perspective are challenged. In threatening the intellectual perspective while leaving the motivational perspective unchallenged, akratic action poses apparent difficulties for the project of combining the two perspectives into a unified outlook on the explanation of intentional human action. That is a primary source of perennial philosophical interest in akratic action.

There is much to recommend the motivational and intellectual perspectives, and a plausible combination is theoretically desirable. To some theorists, the threat that strict akratic action poses to a unified, motivational/intellectual perspective seems so severe that they deem such action conceptually or psychologically impossible (Hare 1963 , chap. 5; Pugmire 1982 ; Watson 1977). [6] Many others try to

end p.242

accommodate strict akratic action in a unified perspective (Davidson 1980 , chap. 2, 1982 ; Dunn 1987 ; Mele 1987 , 1995 ; Pears 1984).

2. Explaining Strict Akratic Action

To the extent that one's decisive judgment derives from one's motivational attitudes, it has a motivational dimension. [7] That helps explain why many regard akratic action as theoretically perplexing. How, they wonder, can the motivation associated with a judgment of this kind be outweighed by competing motivation, especially when the competing motivational attitudes—or *desires,* broadly construed—have been taken into account in arriving at the judgment?

Elsewhere (Mele 1987) I defended an answer to this question that rests partly on two theses, both of which I defended.

1. Decisive judgments normally are formed at least partly on the basis of our evaluation of the "objects" of our desires (i.e., the desired items).
2. The motivational force of our desires does not always match our evaluation of the objects of our desires. (Santas 1966 , Smith 1992 , Stocker 1979 , Watson 1977) [8]

If both theses are true, it should be unsurprising that sometimes, although we decisively judge it better to *A* than to *B,* we are more strongly motivated to *B* than to *A.* Given how our motivation stacks up, it should also be unsurprising that we *B* rather than *A.*

Thesis 1 is a major plank in a standard conception of practical reasoning. In general, when we reason about what to do, we inquire about what it would be best, or better, or "good enough" to do, not about what we are most strongly motivated to do. When we ask such questions while having conflicting desires, our answers typically rest significantly on our assessments of the objects of our desires—which may be out of line with the motivational force of those desires, if thesis 2 is true.

Thesis 2, as I argued in Mele 1987 , is confirmed by common experience and thought experiments and has a foundation in empirical studies. Desire-strength is influenced not only by our evaluation of the objects of desires, but also by such factors as the perceived proximity of prospects for desire-satisfaction, the salience of desired objects in perception or in imagination, and the way we attend to desired objects (Ainslie 1992 , Metcalfe and Mischel 1999 , Rorty 1980a). Factors such as these need not have a matching effect on assessment of desired objects.

A few hours ago, an agent decisively judged it better to *A*than to *B,*but he

end p.243

now has a stronger desire to *B*than to *A.*Two versions of the case merit attention. In one, along with the change in desire strength, there is a change of judgment. For example, last night, after much soul-searching, Al formed a decisive judgment favoring not eating after-dinner snacks for the rest of the month and desired more strongly to forego them than to indulge himself; but now, a few hours after dinner, Al's desire for a snack is stronger than his desire for the rewards associated with not snacking, and he decisively judges it better to have a snack than to refrain. In another version of the case, the change in relative desire strength is not accompanied by a change of judgment. Al retains the decisive judgment favoring not eating after dinner, but he eats anyway. Assuming that Al eats intentionally and is not compelled to eat, this is a strict akratic action.

Empirical studies of the role of representations of desired objects in impulsive behavior and delay of gratification (reviewed in Mele 1987 , 88–93; see Mischel et al. 1989 for an overview) provide ample evidence that our representations of desired objects have two important dimensions, a motivational and an informational one. Our decisive judgments may be more sensitive to the informational dimension of our representations than to the motivational dimension, with the result that such judgments sometimes recommend courses of action that are out of line with what we are most strongly motivated to do at the time. If so, strict akratic action is a real possibility—provided that at least some intentional actions that conflict with agents' decisive judgments at the time of action are not *compelled.*

A discussion of compulsion would lead quickly to the issue of free will, which is well beyond the scope of this chapter. It is worth noting, however, that unless a desire is irresistible, it is up to the agent, in some sense, whether she acts on it. This idea is an element of both the motivational and the intellectual perspective on intentional action. Another element is the idea that relatively few desires are irresistible. Of course, a proper appreciation of the latter idea would require an analysis of irresistible desire. [9] It may suffice for present purposes to suggest that, often, when we act against our decisive judgments, we could have used our resources for self-control in effectively resisting temptation. [10] Normal agents can influence the strength of their desires in a wide variety of ways (Ainslie 1992 , Metcalfe and Mischel 1999 , Mischel et al. 1989). For example, they can refuse to focus their attention on the attractive aspects of a tempting course of action and concentrate instead on what is to be accomplished by acting as they judge best. They can attempt to augment their motivation for performing the action judged best by promising themselves rewards for doing so. They can picture a desired item as something unattractive—for example, a chocolate pie as a plate of chocolate-coated chewing

tobacco—or as something that simply is not arousing. Desires typically do not have immutable strengths, and the plasticity of motivational strength is presupposed by standard conceptions of self-control. Occasionally we *do not*act as we judge best, but it is implausible that, in all such cases, we *cannot*
end p.244

act in accordance with these judgments. (This suggestion is defended in Mele 1987 , chap. 2; 1995 , chap. 3; and 2002 .) [11]

3. Kinds of Akratic Action and the Irrationality of Strict Akratic Action

Not all akratic action is of the strict kind. In this section, without aspiring to be exhaustive, I identify two additional kinds discussed in the literature. I also comment on the question of whether strict akratic action is necessarily irrational.
Socrates, in defending the thesis that no one *knowingly*does wrong, argues that what actually happens in apparent instances of strict akratic action is that, owing to the proximity of anticipated pleasures, agents change their minds about what it would be best to do (Plato 1953 355d–357d). Even if he mistakenly denies the reality of strict akratic action, Socrates identifies an important phenomenon. Some such changes of mind are *motivationally biased*processes (on motivated bias, see secs. 4–6). When one is tempted to do something that conflicts with one's decisive judgment, one has motivation to believe that the tempting option would be best. After all, acquiring that belief would diminish one's resistance to acting as one is tempted to act (Pears 1984 , 12–13). In what I elsewhere (Mele 1996) called “Socratic akratic action” (without meaning to suggest that Socrates regarded the episodes as akratic), the agent's new belief issues from a process biased by a desire for the tempting option. In the context of *akrasia,*the most relevant standards for determining bias are the agent's. Perhaps the agent accepts a principle about beliefs that is violated by his present change of mind—for example, the principle that it is best not to allow what one wants to be the case to shape what one believes is the case. And if a motivationally biased change of mind of this kind is avoidable by the agent by means of an exercise of self-control, it is itself an *akratic*episode, an episode manifesting *akrasia*or an associated imperfection, for no perfectly self-controlled person makes motivated judgments that are biased relative to his own standards, if he can avoid doing so by exercising self-control. Furthermore, an intentional action that accords with the new judgment is derivatively akratic (Mele 1987 , 6–7, 1996 ; Pears 1984 , 12–13; Rorty 1980b).
Seemingly, there also are “unorthodox” instances of akratic action, in which agents act *in accordance with*their decisive judgments, and, similarly, unorthodox exercises of self-control in support of conduct that conflicts with the agents' decisive judgments (Bigelow, Dodds, and Pargetter 1990 , 46; Hill 1986 , 112; Jackson
end p.245

1984 , 14; Kennett 2000 , 120–24; Mele 1987 , 7–8, 1995 , 60–76). Here is an illustration of unorthodox akratic action from Mele 1995 : “Young Bruce has decided to join some wayward Cub Scouts in breaking into a neighbor's house, even though he decisively judges it best not to do so.At the last minute, Bruce refuses to enter the house and leaves the scene of the crime. His doing so because his decisive judgment has prevailed is one thing; his refusing to break in owing simply to a failure of nerve is another. In the latter event, Bruce arguably has exhibited weakness of will: he `chickened out'” (60). If, instead, Bruce had mastered his fear and participated in the crime, we would have an unorthodox exercise of self-control. John Bigelow, Susan Dodds, and Robert Pargetter (1990) regard unorthodox episodes of these kinds as support for the idea that what is essential to akratic action is the presence of a second-order desire that either loses or wins against a first-order desire in the determination of action, independently of what (if anything) the agent judges it best to do. [12] For argumentation to the contrary, see Mele 1995 , chap. 4.

I suggested that an agent who acts akratically against a rationally formed decisive judgment acts in a subjectively irrational way. But suppose his judgment is irrational or formed on the basis of reflection that does not take into account important, relevant attitudes of his. And suppose he acts on the basis of attitudes that constitute or reflect better reasons than the ones that ground his judgment. In that case, some have argued, his akratic action is rational, or less irrational than an action in accordance with his decisive judgment would have been (Arpaly 2000 , Audi 1990 , McIntyre 1990). In such cases, an akratic action may reflect an agent's system of values better than his decisive judgment does. [13] However, the fact that a certain akratic action against a particular rationally formed decisive judgment is more coherent with the agent's system of values or reasons than an action in accordance with that judgment would be is consistent with the akratic action's being subjectively irrational, to some degree, in virtue of failing to cohere with the judgment. (The same may be said of akratic actions against decisive judgments that are irrationally formed.) Also, it may be doubted that the problem with a rationally formed decisive judgment that is not sensitive to certain of the agent's values or reasons is irrationality.

4. Motivated Irrational Belief: Agency and Anti-agency Views

“Biased” or “irrational” beliefs are biased or irrational relative to some standard or other. In the context of akratic belief (Davidson 1985 ; Heil 1984 ; Mele 1987 , chap. 8; Pears 1984 , chap. 4; Rorty 1983), the germane standards are the believer's own. In work on self-deception, general epistemic standards are typically assumed. One test for motivationally biased belief of a sort appropriate to self-deception is the following. If *S*is self-deceived in believing that *p,*and *D*is the collection of relevant data readily available to *S,*then if *D*were made readily available to *S*'s impartial cognitive peers (including merely hypothetical people) and they were to engage in at least as much reflection on the issue as *S*does and at least a moderate amount of reflection, those who conclude that *p*is false would significantly outnumber those who conclude that *p*is true (cf. Mele 2001 , 106). Two plausible requirements for impartiality in this context are that one neither

desire that *p* nor desire that ~*p* and that one not prefer avoidance of either of the following errors over the other: falsely believing that *p* and falsely believing that ~*p*.

The question whether all motivationally biased beliefs are irrational raises an intriguing question about rationality. Might a motivationally biased belief be rational—or, at least, not irrational—from some legitimate point of view? W. K. Clifford asserts that "it is wrong always, everywhere, and for anyone, to believe anything upon insufficient evidence" (1886 , 346). William James denies this, contending that "a rule of thinking which would absolutely prevent me from acknowledging certain kinds of truth if those kinds of truth were really there, would be an irrational rule" ([1897] 1979 , 31–32). It is at least consistent with James's claim that a person who is rightly convinced that he would be miserable if he were no longer to believe that God exists may rationally believe, on insufficient evidence, that God exists. I set this question about rationality aside and pursue the issue of motivationally biased beliefs. These beliefs are irrational by general epistemic standards or from an epistemic point of view.

Consider two bold theses about motivationally biased beliefs:

1. *The agency view.* In every instance of motivationally biased belief that *p,* we try to bring it about that we acquire or retain the belief that *p* or try to make it easier for ourselves to acquire or retain it.
2. *The anti-agency view.* In no instance of motivationally biased belief that *p* does one try to bring it about that one acquires or retains the belief or try to make it easier for oneself to acquire or retain it.

Probably, the truth lies somewhere between these poles. But which thesis is closer to the truth?

One problem for the agency view is central to a familiar puzzle about self-deception. The attempts to which the view appeals threaten to undermine themselves. If Al is trying to bring it about that he believes that he is a good day trader—not by improving his investment skills, but by ignoring or downplaying evidence that he is an inferior trader while searching for evidence that he has superior investment skills—won't he see that the "grounds" for belief that he arrives at in this way are illegitimate? And won't he consequently fail in his at

end p.247

tempt? A predictable reply is that the "tryings" or efforts at issue are not conscious efforts and therefore need not block their own success in this way. [14] Whether, and to what extent, we should postulate unconscious tryings in attempting to explain motivationally biased belief depends on what the alternatives are.

The main problem for the anti-agency view also is linked to this puzzle about self-deception. Apparently, we encounter difficulties in trying to understand how motivationally biased beliefs—or many such beliefs—can arise, if not through efforts of the kind the agency view postulates. How, for example, can Al's wanting it to be the case that he is a good day trader motivate him to believe that he is good at this except by

motivating him to try to bring it about that he believes this or to try to make it easier for himself to believe this? [15] The anti-agency view is faced with a clear challenge: to provide an alternative account of the mechanism(s) by which desires lead to motivationally biased beliefs.

The following remarks by David Pears and Donald Davidson on the self-deceptive acquisition of a motivationally biased belief are concise expressions of two different "agency" views of the phenomenon:

[There is a] sub-systembuilt around the nucleus of the wish for the irrational belief and it is organized like a person. Although it is a separate centre of agency within the whole person, it is, from its own point of view, entirely rational. It wants the main system to form the irrational belief and it is aware that it will not form it, if the cautionary belief [i.e., the belief that it would be irrational to form the desired belief] is allowed to intervene. So with perfect rationality it stops its intervention. (Pears 1984 , 87)

His practical reasoning is straightforward. Other things being equal, it is better to avoid pain; believing he will fail the exam is painful; therefore (other things being equal) it is better to avoid believing he will fail the exam. Since it is a condition of his problem that he take the exam, this means it would be better to believe he will pass. He does things to promote this belief. (Davidson 1985b , 145–46)

Both views rest mainly on the thought that the only, or best, way to account for certain data is to hold that the person, or some center of agency within her, tries to bring it about that she, or some "system" in her, holds a certain belief.

Consider a case of self-deception similar to the one Davidson diagnoses in the passage last quoted. Carlos "has good reason to believe" that he will fail his driver's test (Davidson 1985b , 145). "He has failedtwiceand his instructor has said discouraging things. On the other hand, he knows the examiner personally, and he has faith in his own charm.The thought of failingagain is painful" (145–46). Suppose the overwhelming majority of Carlos's impartial cognitive peers presented with his evidence would believe that Carlos will fail and none would believe that he will pass. (Some peers with high standards for belief may withhold belief.) Even so, in the face of the contrary evidence, Carlos believes that he will pass. Predictably, he fails.

end p.248

Self-deception is often thought to be such that if Carlos is self-deceived in believing that he will pass the test, he believed at some time that he would fail it (Bach 1981 , Demos 1960 , Haight 1980 , Quattrone and Tversky 1984 , Rorty 1972 , Sackeim and Gur 1985). However, in accommodating the data offered about Carlos, there is no evident need to suppose he had this true belief. Perhaps his self-deception is such that not only does he acquire the belief that he will pass, but he never acquires the belief that he will fail. In fact, this seems true of much self-deception. Seemingly, some parents who are self-deceived in believing that their children have never experimented with drugs and some people who are self-deceived in believing that their spouses have not had affairs have never believed that these things have happened. Owing to self-deception, they have not come to believe the truth, and perhaps they never will.

That having been said, there probably are cases in which a person who once believed an unpleasant truth, p, later is self-deceived in believing that $\sim p$. For example, a mother who

once believed that her son was using drugs subsequently comes to believe that he has never used drugs and is self-deceived in so believing. Does a change of mind of this sort *require*an exercise of agency of the kind postulated by Pears or Davidson? Is such a change of mind *most plausibly explained*, at least, on the hypothesis that some such exercise of agency occurred? A theorist who attends to the stark descriptions Pears and Davidson offer of the place of agency in self-deception should at least wonder whether things are so straightforward.

It is often held that, in Jeffrey Foss's words, "desires have no explanatory force without associated beliefs" that identify (apparent) means to the desires' satisfaction, and this is part of "thelogic of belief-desire explanation" (1997 , 112). This claim fares poorly in the case of motivationally biased belief. "A survey of one million high school seniors found that25% thought they were in the top 1%" in ability to get along with others (Gilovich 1991 , 77). A likely hypothesis about this striking figure includes the idea that desires that *p*can contribute to biased beliefs that *p*. If Foss's claim were true, a student's wanting it to be true that she is exceptionally affable would help explain her believing that she is only in combination with an instrumental belief (or a collection thereof) that links her believing that she is superior in this sphere to the satisfaction of her desire to be superior. But we search in vain for instrumental beliefs that are plausibly regarded as turning the trick frequently enough to accommodate the data. Perhaps believing that one is exceptionally affable can help bring it about that one is superior in this sphere, and some high school students may believe that this is so. But it is highly unlikely that most people who have a motivationally biased belief that they are exceptionally affable have that belief, in part, *because*they want it to be true that they are superior in this area *and*believe that believing that they are superior can make it so. No other instrumental beliefs look more promising.

Should we infer, then, that wanting it to be true that one has a superior

end p.249

ability to get along with others plays a role in explaining only relatively few unwarranted beliefs that one is superior in this area? Not at all. There is powerful empirical evidence, some of which is reviewed shortly, that desiring that *p*makes a broad causal contribution to the acquisition and retention of unwarranted beliefs that *p*. Desires that do this properly enter into causal *explanations*of the pertinent biased beliefs. It is a mistake to assume that the role characteristic of desires in explaining intentional actions is the only explanatory role they can have.

5. Evidence for and Sources of Motivationally Biased Belief

If Pears or Davidson is right about a case like the mother's or Carlos's, presumably similar exercises of agency are at work in an enormous number of high school students who believe that, regarding ability to get along with others, they are "in the top 1%." Perhaps self-deception is very common, but the same is unlikely to be true of intentional self-manipulation of the kind Pears or Davidson describes. Theorists inclined to agree

with Foss's claim about the explanatory force of desires will be inclined toward some version or other of the agency view of motivationally biased belief and self-deception. However, as I will explain, desires contribute to the production of motivationally biased beliefs, including beliefs that one is self-deceived in holding, in a variety of relatively well understood ways that fit the anti-agency model.

The survey I mentioned also found that 70 percent of high school seniors "thought they were above average in leadership ability, and only 2% thought they were below average." "A survey of university professors found that 94% thought they were better at their jobs than their average colleague" (Gilovich 1991 , 77). Data such as these suggest that desires sometimes bias beliefs. The aggregated self-assessments are wildly out of line with the facts (e.g., only 1 percent can be in the top 1 percent), and the qualities asked about are desirable. There is powerful evidence that people have a tendency to believe propositions that they want to be true even when an impartial investigation of readily available data would indicate that they probably are false. A plausible hypothesis about that tendency is that desire sometimes biases belief.

Controlled studies provide confirmation for this hypothesis. In one study, 75 women and 86 men read an article asserting that "women were endangered by caffeine and were strongly advised to avoid caffeine in any form"; that the major danger was fibrocystic disease, "associated in its advanced stages with breast can

cer"; and that "caffeine induced the disease by increasing the concentration of a substance called cAMP in the breast" (Kunda 1987 , 642). (Because the article did not directly threaten men, they were used as a control group.) Subjects were then asked to indicate, among other things, "how convinced they were of the connection between caffeine and fibrocystic disease and of the connection between caffeine andcAMP on a 6-point scale" (643–44). Female "heavy consumers" of caffeine were significantly less convinced of the connections than female "low consumers." The males were considerably more convinced than the female "heavy consumers"; and there was a much smaller difference in conviction between "heavy" and "low" male caffeine consumers (the heavy consumers were slightly *more*convinced of the connections). Because all subjects were exposed to the same information and the female "heavy consumers" were the most seriously threatened by it, a plausible hypothesis is that a desire that their coffee drinking has not significantly endangered their health helps to account for their lower level of conviction (Kunda 1987 , 644). Indeed, in a study in which the reported hazards of caffeine use were relatively modest, "female heavy consumers were no less convinced by the evidence than were female low consumers." Along with the lesser threat, there is less motivation for skepticism about the evidence.

Attention to some phenomena that have been argued to be sources of *unmotivated*biased belief sheds light on motivationally biased belief. A number of such sources have been identified, including the following two.

1. *Vividness of information*. A datum's vividness for us often is a function of such things as its concreteness, its "imagery-provoking" power, and its sensory, temporal, or spatial proximity (Nisbett and Ross 1980 , 45). Vivid data are more likely to be recognized, attended to, and recalled than pallid data. Consequently, vivid data tend to have a disproportional influence on the formation and retention of beliefs.
2. *The confirmation bias*. People testing a hypothesis tend to search (in memory and the world) more often for confirming than for disconfirming instances and to recognize the former more readily (Baron 1988 , 259–65; Klayman and Ha 1987 ; Nisbett and Ross 1980 , 181–82). This is true even when the hypothesis is only a tentative one (as opposed, e.g., to a belief one has). People also tend to interpret relatively neutral data as supporting a hypothesis they are testing (Trope, Gervey, and Liberman 1997 , 115).

Although sources of biased belief apparently can function independently of motivation, they also may be triggered and sustained by desires in the production of *motivationally*biased beliefs. [16] For example, desires can enhance the vividness or salience of data. Data that count in favor of the truth of a proposition that one hopes is true may be rendered more vivid or salient by one's recognition that they so count. Similarly, desires can influence which hypotheses occur to one and affect the salience of available hypotheses, thereby setting the stage for the confirmation bias. [17] Owing to a desire that *p*, one may test the hypothesis that *p*is true rather than the contrary hypothesis. In these ways and others, a desire that *p*may help explain the acquisition of an unwarranted belief that *p*.

Sometimes we generate our own hypotheses, and sometimes others suggest hypotheses to us—including extremely unpleasant ones. If we were always to concentrate primarily on confirmation in hypothesis testing, independently of what is at stake, that would indicate the presence of a cognitive tendency or disposition that uniformly operates independently of desires and that desires never play a role in influencing the proportion of attention we give to evidence for the falsity of a hypothesis. However, there is powerful evidence that the "confirmation bias" is much less rigid than this. For example, in one study (Gigerenzer and Hug 1992), two groups of subjects were asked to test "social-contract rules such as If someone stays overnight in the cabin, then that person must bring along a bundle of firewood'" (Friedrich 1993 , 313). The group asked to adopt "the perspective of a cabin guard monitoring compliance" showed an "extremely high frequency" of testing for disconfirmation (i.e., for visitors who stay in the cabin overnight but bring no wood). The other group, asked to "take the perspective of a visitor trying to determine" whether firewood was supplied by visitors or by a local club, displayed the common confirmation bias. [18]

6. A Motivational Model of Lay Hypothesis Testing

An interesting recent theory of lay hypothesis testing is designed, in part, to accommodate data of the sort I have been describing. I explored it in Mele 2001 , where I offered grounds for caution and moderation and argued that a qualified version is

plausible. [19] I named it the "FTL theory," after the authors of the two essays on which I primarily drew, Friedrich 1993 and Trope and Liberman 1996 . Here, I offer a thumbnail sketch.

The basic idea of the FTL theory is that a concern to minimize costly errors drives lay hypothesis testing. The *error*son which the theory focuses are false beliefs. The *cost*of a false belief is the cost, including missed opportunities for gains, that it would be reasonable for the person to expect the belief—if false—to have, given his desires and beliefs, if he were to have expectations about such things. A central element of the FTL theory is a "confidence threshold"—or a "threshold," for short. The lower the threshold, the thinner the evidence sufficient

end p.252

for reaching it. Two thresholds are relevant to each hypothesis: "The acceptance threshold is the minimum confidence in the truth of a hypothesis," *p*, sufficient for acquiring a belief that *p*"rather than continuing to test [the hypothesis], and the rejection threshold is the minimum confidence in the untruth of a hypothesis," *p*, sufficient for acquiring a belief that *p*"and discontinuing the test" (Trope and Liberman 1996 , 253). The two thresholds often are not equally demanding, and acceptance and rejection thresholds respectively depend "primarily" on "the cost of false acceptance relative to the cost of information" and "the cost of false rejection relative to the cost of information." The "cost of information" is simply the "resources and effort" required for gathering and processing "hypothesis-relevant information" (252).

Confidence thresholds are determined by the strength of aversions to specific costly errors together with information costs. Setting aside the latter, the stronger one's aversion to falsely believing that *p*, the higher one's threshold for belief that *p*. These aversions influence belief in a pair of related ways. First, because, other things being equal, lower thresholds are easier to reach than higher ones, belief that ~*p*is a more likely outcome than belief that *p*, other things being equal, in a hypothesis tester who has a higher acceptance threshold for *p*than for ~*p*. Second, the aversions influence *how*we test hypotheses, not just *when we stop*testing them (owing to our having reached a relevant threshold). Recall the study in which subjects asked to adopt "the perspective of a cabin guard" showed an "extremely high frequency" of testing for disconfirmation, whereas subjects asked to "take the perspective of a visitor" showed the common confirmation bias.

It might be claimed that if aversions to specific errors function in the second way just identified, they work together with beliefs to the effect that testing-behavior of a particular kind is conducive to avoiding these errors. It might be claimed, accordingly, that the pertinent testing-behavior is performed with the intention of avoiding, or of trying to avoid, the pertinent error. The thrust of these claims is that the FTL theory accommodates the confirmation bias, for example, by invoking a model of intentional action.

This is not a feature of the FTL model, as its proponents understand it. Friedrich, for example, claims that desires to avoid specific errors can trigger and sustain "automatic test strategies" (313), which supposedly happens in roughly the nonintentional way in which a desire that *p*results in the enhanced vividness of evidence for *p*. In Mele 2001

(41–49, 61–67), I argued that a person's being more strongly averse to falsely believing that ~*p* than to falsely believing that *p* may have the effect that he primarily seeks evidence for *p*, is more attentive to such evidence than to evidence that ~*p*, and interprets relatively neutral data as supporting *p*, without this effect's being mediated by a belief that such behavior is conducive to avoiding the former error. The stronger aversion may simply frame the topic in such a way as to trigger and sustain these manifestations of the confirmation bias without the assistance of a belief that behavior of this kind is
end p.253

a means of avoiding particular errors. Similarly, having a stronger aversion that runs in the opposite direction may result in a skeptical approach to hypothesis testing that in no way depends on a belief to the effect that an approach of this kind will increase the probability of avoiding the costlier error. Given the aversion, skeptical testing is predictable independently of the agent's believing that a particular testing style will decrease the probability of making a certain error.

The FTL theory applies straightforwardly to both "straight" and "twisted" self-deception (Mele 2001 , 4–5, 94–118). In straight cases, we are self-deceived in believing something that we want to be true. In twisted cases, we are self-deceived in believing something that we want to be *false* (and do not also want to be true). Twisted self-deception may be exemplified by an insecure, jealous husband who believes that his wife is having an affair despite possessing only relatively flimsy evidence for that proposition and despite unambivalently wanting it to be false that she is so engaged. [20] Friedrich writes: "A prime candidate for primary error of concern is believing as true something that leads [one] to mistakenly criticize [oneself] or lower [one's] self-esteem. Such costs are generally highly salient and are paid for immediately in terms of psychological discomfort. When there are few costs associated with errors of self-deception (incorrectly preserving or enhancing one's self-image), mistakenly revising one's self-image downward or failing to boost it appropriately should be the focal error" (314). Here, he plainly has straight self-deception in mind.

Whereas, for many people, it may be more important to avoid acquiring the false belief that their spouses are having affairs than to avoid acquiring the false belief that they are not so engaged, the converse may well be true of some insecure, jealous people. The belief that one's spouse is unfaithful tends to cause significant psychological discomfort. Even so, avoiding falsely believing that their spouses are faithful may be so important to some people that they test relevant hypotheses in ways that, other things being equal, are less likely to lead to a false belief in their spouses' fidelity than to a false belief in their spouses' infidelity. Furthermore, data suggestive of infidelity may be especially salient for these people and contrary data quite pallid by comparison. Don Sharpsteen and Lee Kirkpatrick observe that "the jealousy complex"—that is, "the thoughts, feelings, and behavior typically associated with jealousy episodes"—is interpretable as a mechanism "for maintaining close relationships" and appears to be "triggered by separation, or the threat of separation, from attachment figures" (1997 , 627). It certainly is conceivable that, given a certain psychological profile, a strong desire to maintain one's relationship with one's spouse plays a role in rendering the potential error of falsely believing one's spouse to be innocent of infidelity a "costly" error, in the FTL sense, and more costly

than the error of falsely believing one's spouse to be guilty. After all, the former error may reduce the probability that one takes steps to protect the relationship against an intruder. The FTL theory
end p.254

provides a basis for a plausible account of twisted self-deception (Mele 2001 , chap. 5).

7. Conclusion: The Paradox of Irrationality

Donald Davidson writes: “The underlying paradox of irrationality, from which no theory can entirely escape, is this: if we explain it too well, we turn it into a concealed form of rationality; while if we assign incoherence too glibly, we merely compromise our ability to diagnose irrationality by withdrawing the background of rationality needed to justify any diagnosis at all” (1982 , 303). The explanations sketched here of strict akratic action and motivationally biased belief avoid Davidson's worries about paradox. Akratic agents act for reasons, and in central cases, they make rational decisive judgments: “the background of rationality” required for that is in place. But insofar as their uncompelled actions are at odds with their rational decisive judgments, they act irrationally. Motivationally biased believers test hypotheses and believe on the basis of evidence. Again there is a background of rationality. But, owing to the influence of motivation, they violate general standards of epistemic rationality.

NOTES

Parts of this chapter derive from Mele 1987 , 1995 , 1998 , and 2001. I am grateful to Piers Rawling for comments on a draft.

1. For a detailed account, see Mele 1995 , chaps. 1–7.
2. Assuming a middle ground between *akrasia* and self-control, not all akratic actions manifest *akrasia*. Someone who is more self-controlled than most people in a certain sphere may, in a particularly trying situation, succumb to temptation in that sphere against her better judgment. If her intentional action is uncompelled, she has acted akratically—even if her action manifests not *akrasia* but an associated imperfection.
3. An agent who makes such a judgment may or may not proceed to search for additional options. He may regard the best member of his currently envisioned options as “good enough.”
4. On failures of coherence, see Arpaly 2000 and Harman, chap. 3, this volume.
5. On the nature of motivational strength and the theoretical utility of the notion, see Mele 2003a , chap. 7.
end p.255

6. For replies to Hare, Pugmire, and Watson, see Mele 1987 , chap. 2 and 51–55. See also Pugmire 1994 , responding to Mele 1987 , and the rejoinder in Mele 1995 , 44–54.
7. This is not to say that motivation is “built into” the judgment itself.

8. For opposition to the idea that desires vary in motivational strength, see Charlton 1988 , Gosling 1990 , and Thalberg 1985 . For a reply to the opposition, see Mele 2003a , chap. 7.
9. See Mele 1992 , chap. 5, for an analysis of irresistible desire.
10. This is not a necessary condition for strict akratic action. There are Frankfurt-style cases (Frankfurt 1969) in which, although one *A*-ed akratically and without any external interference, if one had been about to resist temptation, a mind-reading demon would have prevented one from doing so (see Mele 1995 , 94–95).
11. One might claim that anyone who is more strongly motivated to *B* than to *A* will also be more strongly motivated to allow that feature of her motivational condition to persist than to change it. I rebut this claim in Mele 1987 , chap. 6, and Mele 1995 , chap. 3.
12. For similar positions on akratic action, see Schiffer 1976 and Swanton 1992 , chap. 10. Swanton contends that "in the context of weakness of will, the will should be identified with strong evaluation," a certain kind of "evaluative second-order desire" (149). Also see Jeffrey 1974 . On an alternative view, the agent who akratically *A*-s believes that she should believe that it is best not to *A* but does not believe what she believes she should (Tenenbaum 1999 ; cf. Buss 1997 , 36).
13. Dion Scott-Kakures (1997) argues that akratic agents are wrong about what they have "more reason" to do.
14. See, e.g., Bermudez 2000 , Quattrone and Tversky 1984 , Sackeim 1988 , and Talbott 1995 . A related response is mental partitioning: the deceived part of the mind is unaware of what the deceiving part is up to. See Pears 1984 (cf. 1991) for a detailed response of this kind and Davidson 1985 (cf. 1982) for more modest partitioning. For criticism of partitioning views of self-deception, see Barnes 1997 , Johnston 1988 , and Mele 1987 .
15. Two uses of "motivate" should be distinguished. In one, a desire's motivating an action or a belief is a matter of its *being* motivation for it. Piers's desire to fish today is motivation for him to fish, even if, desiring more to work on his chapter, he foregoes a fishing trip. In another use, a desire motivates something only if, in addition to being motivation for it, it plays a role in *producing* that thing. Here, I use "motivate" in the second sense.
16. I develop this idea in Mele 1987 , chap. 10, and Mele 2001 . Kunda 1990 develops the same theme, concentrating on evidence that motivation sometimes primes the confirmation bias. Also see Kunda 1999 , chap. 6.
17. For motivational interpretations of the confirmation bias, see Friedrich 1993 and Trope and Liberman 1996 , 252–65.
18. For further discussion, see Samuels and Stich, chap. 15, this volume.
19. See Mele 2001 , 31–49, 63–70, 90–91, 96–98, 112–18.
20. On this case, see Barnes 1997 , chap. 3; Lazar 1999 , 274–77; and Pears 1984 , 42–44. Also see Davidson 1985 , 144; Demos 1960 , 589; McLaughlin 1988 , 40; Mele 1987 , 114–18; and Mele 2001 , chap. 5.

chapter 14 PARADOXES OF RATIONALITY

Roy Sorensen

Fear my dot. That's right, the little round mark I used as a period. Having trouble? Would a free copy of this book help?

1. The Scope of Reasons

The book bonus gives you a reason to fear the little dot. But this reason does not make fear of my dot rational. The reason is at the wrong level. You would rationally fear my dot if physicists had just persuaded you it was a black hole. A reason can make the fear rational only by conforming to the inner logic of fear—by showing that the object is dangerous. A reason to *fear* that the dot will harm you need not be a reason to fear *that the dot will harm you* (Sorensen 1998).
In *Julius Caesar*, Shakespeare suggests that fear of death is irrational because the emotion makes one miserable without the prospect of a compensating benefit:
Cowards die many times before their deaths;
The valiant never taste of death but once.
Of all the wonders that I yet have heard,
end p.257

It seems to me most strange that men should fear;
Seeing that death, a necessary end,
Will come when it will come. (II.ii.37–42)
Let us charitably interpret the poem as only a criticism of fear of death itself (and not fear of a premature death). We who cower before the Great Scythe should then concede that Shakespeare has given us a reason to suppress our fear of death. But is that enough to show that fear of death is irrational?
Consider a man out on a ledge. He must escape a burning building by walking a long plank to a neighboring building. There is a real danger of falling, so the inner logic of fear is satisfied. But this fear is apt to cause a misstep. Accordingly, the man stifles his fear. (And his fear of the fear.) He averts his eyes from the abyss. He may even cultivate defiant anger against the fire to make his steps resolute. This anger is irrational. But out on the ledge, it is rational to cultivate a useful irrational emotion. (Fear of my dot may be like that.) And out on the ledge, it is rational to suppress a rational fear. (Fear of death may be like that.)

2. Emotions toward Fictions

The audience watching *Jaws* shrieks. They seem to fear the giant shark even though they are high and dry.
We shriekers know that the shark does not even exist. This leads many aestheticians to conclude that the audience is experiencing only make-believe fear (Walton 1990). Others

say that fear does not always require belief that is one in danger. They think that the mere thought of danger is enough for fear (Yanal 1999).
These theorists are afraid of Colin Radford's (1975) conclusion that the members of the audience are inconsistent. Radford says members of the audience believe that they are in danger and believe that they are in no danger. This would explain why they do not flee the theater and yet are more reluctant to swim in the ocean. The young director of *Jaws,* Stephen Spielberg, had included a scene in which a boy is eaten by the shark. At a test screening, this spectacle caused a man in the audience to proceed to the restroom and vomit. Spielberg became distraught, fearing that his movie was "over the top." But then the purged patron returned to his seat. Spielberg concluded that *Jaws* would be successful.
Spielberg could agree with Radford that those who fear the shark in *Jaws* are being irrationally inconsistent. But Spielberg is still free to approve of the fear. Ditto for consumers of horror movies who rationally pay for the thrill of irrational
end p.258

fears. Those who insist on consistency must watch *Jaws* under the deflating reminder that "it is only a movie."

3. The Range of the Paradoxes

Consistency is central to *theoretical* rationality. When I saw *Jaws,* I believed that the shark's first victim was a beautiful woman. I believed the woman died on July 1 because the sheriff reports she was found on July 2 and the autopsy indicated she had been dead one day. I believed a boy died shortly after, on June 29, which is the date listed on a reward poster. When I was describing the plot of *Jaws,* someone pointed out my inconsistent chronology. Argh! Red-faced, I withdrew my beliefs about the dates of the shark attacks.
The pain of contradiction also accounts for the power of paradoxes to change our minds. Promoters of a paradox present a *small* set of propositions that are individually plausible but jointly inconsistent. The importance of the size of the set became apparent with the discovery of the paradox of the preface (Makinson 1965), an author who apologizes for the errors that are bound to be in his text. This belief expressed in the preface is correct if and only if one of the beliefs he expresses in the text is mistaken. Therefore, the modest author is bound to have a false belief. Since it is impossible for all of his beliefs to be true, they are jointly inconsistent. Does this show that it is rationally permissible to have inconsistent beliefs? Well, let's not understate the implication. Since acknowledgment of one's general fallibility seems mandatory, we wonder whether all rational human beings know themselves to be inconsistent.
Practical rationality is broader than theoretical rationality. It encompasses action (as opposed to mere behavior like hiccups). Action is based on desire as well as belief, so anomalous features of preferences will affect practical rationality.

Just as causation is the cement of the physical universe, rationality is the cement of our mental lives. Causation binds each event to preceding events and future events. The result is a conglomerate of events that unifies the subject matter of physics. Rationality binds beliefs and desires into explanations of actions. The actions of a single agent fit together into a coherent whole. Agents are the subject matter of the social scientists and the humanities—and the focus of attention for any healthy human being.

We all have anecdotes illustrating how the postulation of beliefs and desires creates satisfying explanations. As a newcomer to New Hampshire winters, I was puzzled to see cars parked with their windshield wipers pulled away from their windshields in apparent homage to the sky. When I started my car, I found that

end p.259

my wipers were frozen to the windshield. Mystery solved! Now I too point my windshield wipers to the sky when I *believe* that there will be freezing rain because I *want* my windshield wipers mobile.

The social sciences are a continuation of this commonsense practice of rationally reconstructing actions. I ask the economist: Why do so many Mexicans immigrate? He answers: Because the difference in the standard of living between Mexico and the United States is large and the cost of crossing the border is small. Thousands of Mexicans *want* a higher standard of living and *believe* moving to the United States is the most efficient means of satisfying that desire. So Mexicans come in proportion to the gap in the standard of living minus the cost of border crossing.

Belief-desire explanations are also employed by biologists. Why do cheetahs hunt at noon? It is difficult to surprise prey in broad daylight. Isn't it easier and more comfortable to hunt sleeping animals in the cool of the night, under the cover of darkness? Answer: Yes, that's why *lions* hunt at night. Lions chase cheetahs and steal their meals. Cheetahs want to eat what they kill and believe they are more likely to be unmolested while the lions siesta.

Just as folk physics contains factual errors, gaps, and conflicts, folk psychology is marred by myths, incompleteness, and inconsistency. Many paradoxes of rationality are signs of these flaws. In this respect, they are bad news. But awareness of a problem is the first step toward solving it. Optimists welcome the discovery of a paradox as an opportunity to make an improvement over raw common sense. The resulting theory itself becomes the basis of new expectations. Since the theorist's esoteric expectations can themselves be confounded, paradoxes are part of a generic cycle of self-correction.

4. Riddles of Maximization

Many paradoxes orbit the principle of maximizing expected utility. Blaise Pascal formulated this principle in his seminal studies of gambling. Rational agents consider both the value of the outcome and the probability of that outcome. Since probabilities can be measured on a scale between 0 and 1, we can multiply probabilities with values to obtain expected values.

Pascal recognized that value can be measured in various ways. The egoist cares only for his well-being, the altruist cares for all. The hedonist Jeremy Bentham was later to assert that moral rightness was merely a matter of maximizing pleasure and minimizing pain. Thus Pascal's principle of maximizing expected value came to be incorporated into utilitarianism.
end p.260

Pascal thought his theory of gambling afforded a new basis for theism. Whereas most arguments for believing in God's existence try to prove God exists, Pascal tries to prove only that it is prudent to believe that God exists. If God sends believers to heaven and nonbelievers to hell, then there is an incentive to believe that God exists. The payoff seems infinite. Consequently, any nonzero probability that God exists implies that the expected value of believing in God is infinite. Thus Pascal hoped to bypass epistemological limits and metaphysical deadlocks.
Pascal's Wager has not filled the pews with prudent gamblers. However, it has provoked discussion about how to model infinite value. These speculations have not won the confidence of economists. They concentrate on circumstances involving finite values. Some use the axioms pitched to finite values as premises for the conclusion that value must be finite. This evaluative finitism is less popular with philosophers. Indeed, in the last twenty years there has been a spate of new puzzles about infinite value and some mathematically literate efforts to solve them (Vallentyne and Kagan 1997).
Some infinite decision puzzles involve *finite* goods and infinite time. Before you gleams a bottle of Ever Better wine (a gift from John Pollock [1983]). The wine slowly improves with age. (Assume it asymptotically approaches the quality of a Château Haut-Brion.) More good news: You are immortal. Consequently, you are indifferent as to when you consume a particular good. When should you drink the wine?
Not now. The wine will be better later.
Not later. For at any given time it will be true that the wine will be even better if you waited longer.
But if you do not drink the wine now and do not drink it later, then you will not drink it at all!
What went wrong? You have lived according to the principle that you should not do what you will regret. Whenever you drink the wine as well, you will regret drinking it prematurely. And if you fail to drink the wine, that would be the most regrettable of all your possible alternatives. Should you abandon the prohibition against doing what you will regret?
Michael Slote (1989) thinks Ever Better wine is a counterexample to the principle that one ought not to wittingly choose an inferior alternative. Optimizers say we should choose what is best. But when there is no best alternative, we must resign ourselves to choosing what is good enough. Slote endorses Herbert Simon's (1983) conclusion that rationality is about satisficing rather than optimizing. People have limited resources and so are designed to make decisions that are merely "good enough."
But rationality demands opportunism. Imagine that you are well off but could double your fortune merely by lifting a finger. Is it rationally permissible to forego lifting a finger?

How would Mother Nature answer? “Living things are not selected for their capacity to simply stay alive; they are staying alive in competition with other such living things. The trouble with satisficing as a concept is that it leaves out the competitive element which is fundamental to all life” (Dawkins 1982 , 45–46). Natural selection is sensitive to research costs. So Mother Nature will outfit us with quick and dirty heuristics that make us look like complacent satisficers. But out in the real world, one-upmanship is relentless. If not an optimizer, Mother Nature is a meliorist—always lashing her creatures on to do better. Human beings have acquired metacognitive capacities that enable us to assess how well we are doing. Human beings ask: Will I remember this or should I write it down? Is it likely that I got all these calculations correct or should I double-check? Should I abandon interviewing and just hire on the basis of résumés? When our reasoning seems to go awry, we are apt to detect the malfunction and probe for a fallacy. However, these self-diagnoses are themselves imperfect. Often we do not know whether to blame our intuitions or our principles. Many paradoxes are rooted in this recalcitrant indecision.
Constrained maximizers think the solution to the problem of when to drink Ever Better wine is a *resolution* (McClennen 1990). If you make a resolute choice to drink in ten years, the wine will be better and you will have actually drunk it.
Suppose the ten years have now past. Should you keep your resolution? The unconstrained maximizer answers no: The wine will be better if you wait. The mere fact that one is breaking a resolution is irrelevant. You cannot make something a reason merely by declaring it is a reason (Broome 2001a). Consequently, the unconstrained maximizer says that rational agent is not governed by the dead hand of the past.
Only the future matters! The unconstrained maximizer concedes that this forward-looking feature of rationality is often inconvenient. When in a foreign land with little money, a sleepy traveler would like to stay in a hotel and mail the greedy proprietor her exorbitant fee upon returning to his own country. The gouging proprietor would like that as well. But both know that the traveler would have no reason to mail the exorbitant fee once he returned home. Thus the traveler must spend the night on a park bench.
Punishment raises the same issue. What do you do when your threat fails to deter? Carrying through on the threat is costly to both the punisher and the punished. This cost can be justified when benefits are anticipated from maintaining the credibility of future threats. But sometimes the threatened harm outweighs the value of a steadfast reputation. If the United States had launched a massive first strike against the Soviet Union, then the Soviets would have been without a motive to carry out their threat to retaliate. And vice versa. So how did the policy of mutual assured destruction work? Why was “If you nuke us, I'll nuke you” any more persuasive than the traveler's offer to sleep now and pay later? If the president of the United
end p.262

States knew that he would not counterattack, then he could not intend a counterattack. If the Soviets were to learn that the president was bluffing, then the security of the United States would be jeopardized. (The Soviets were good at detecting bluffers.) Would the president be morally obliged to cultivate a conditional intention to inflict a pointless massacre (Kavka 1978)?

5. Good News or Causal Impact?

Paradoxes have recently stimulated revisions in established theories of rationality. The most famous was precipitated by William Newcomb's problem (Nozick 1970) (table 14.1). You are offered a transparent box and an opaque box. You are free to take either or both. The opaque box might contain a million dollars. It all depends on what a talented psychologist has predicted two weeks ago. If he predicted that you would take both boxes, then he put nothing in the opaque box. If he predicted that you would take only the opaque box, then he put a million dollars in the opaque box. The psychologist is 90 percent accurate. One box or two?

Table 14.1

	Empty opaque box	Full opaque box
Take both boxes	$1000	$1,001,000
Take one box	$0	$1,000,000

On the one hand, taking two boxes dominates taking one box: whatever the contents of the opaque box, you are better off taking both boxes. On the other hand, expected utility calculations suggest that one ought take only one box:
(.9 *x* $1,000,000) > [(.1 *x* $1,000,000) + $1000]
The first edition of Richard Jeffrey's *The Logic of Decision* (1965) implies that the one-box decision is correct. His book articulates an evidentialist theory that counsels us to maximize good news.
Anyone faced with Newcomb's choice should hope that he finds himself walking away with just one box, for the expected value of that box is $900,000. If you find yourself exiting with both boxes, you will be disappointed to learn that the total expected value is only $101,000.
One-boxers and two-boxers agree that you are better off being a one-boxer in a Newcomb situation. But two-boxers believe that this is because Newcomb scenarios reward irrational individuals. When fools are rewarded and sages are penalized, it is best to be a fool. However, this is a truth that can never be practically applied. Weakness of will aside, any action I choose is viewed by me as the best of my available options (even when this option is “looking like a fool”).
Causal decision theorists are two-boxers. They say you should behave in a
end p.263

way that *causes* the best outcome. Producing and sustaining benefits in the world is what counts—not cheering yourself up with statistics.
Since Richard Jeffrey found the two-box solution compelling, he retracted the actuarial aspects of his classic text. His second edition (Jeffrey 1983) accommodates causal decision theory.

Presently, the vast majority of philosophers are two-boxers. Hence most commentators think Newcomb's problem stimulated a fundamental improvement of decision theory.

6. Winning with Irrationality

Jeffrey's revision also affects game theory. In game theory, the outcome depends on your choice plus choices of other people (not just states of nature). For instance, California teenagers play “chicken.” They drive cars toward each other until one of them, the chicken, swerves to avoid a head-on collision.
Losers in these games of brinkmanship are chagrined by how they are *disadvantaged* by their rationality. If persuaded that the other players are irrationally undeterred by the looming mutual disaster, capitulation is rationally mandatory. Little wonder that Nikita Khrushchev would feign irrationality to prevail in Cold War confrontations with the United States. In his classic *The Strategy of Conflict* (1960), Thomas Schelling described a variety of realistic circumstances in which it pays to lose control, to block your acquisition of information, and to engage in other practices that fit the stereotype of irrationality. But these are all cases in which the irrationality is only apparent.
Evolutionary explanations of rationality generally assume that rationality arose because it leads to reproductive success. Cases in which irrationality really is advantageous force qualifications to this account.
end p.264

7. Collective Irrationality from Individual Rationality

The Prisoners' Dilemma (see Bicchieri, chap. 10, and Dreier, chap. 9, this volume) was discovered in 1950 by Merrill Flood and Melvin Dresher while employed by the RAND Corporation. Albert W. Tucker illustrated the game with a tale of two prisoners. The prosecutor has enough evidence to lightly penalize each prisoner. She separately offers leniency to each prisoner in exchange for a confession that could be used against his accomplice. Specifically, she presents row and column with the payoff matrix in table 14.2 .

Table 14.2

	Defect	Do not defect
Defect	3 years, 3 years	1 year, 4 years
Do not defect	4 years, 1 year	2 years, 2 years

As in Newcomb's problem, there is a dominant choice: defecting leaves each prisoner better off regardless of how the other chooses. The resemblance to Newcomb's problem can be strengthened: Suppose that each prisoner knows that his accomplice is so similar

that there is a 90 percent chance that his accomplice will make the same choice as he does. Defecting then yields a worse expected penalty (2.8 years) than not defecting (2.2 years). Game theorists are far more persuaded by the dominance argument for defecting than the expected utility argument for not defecting.

8. The Iterated Prisoners' Dilemma

The case for defecting assumes, for the sake of simplicity, that the prisoners are certain they will not have further contact. If the prisoners think they can penalize each other for defecting, then those consequences must be added to the overall package of sanctions. Since people generally have the prospect of interacting in the future, there is less practical worry about betrayal. In particular, if Prisoners' Dilemma were played repeatedly, one would expect much cooperation.

Flood and Dresher tested this prediction by having two friends play Prisoners' Dilemma one hundred times in sequence (Poundstone 1992 , 106–16). The players usually cooperated (one cooperated 68 times, the other 78 times). The running log each player kept of the reasoning for his move shows that each realized that there was no self-interested motive to cooperate in the last round in the game. Accordingly, each defected on round 100.

But given that each side knows that the other will defect at round 100, neither has a motive to cooperate on round 99. The only point in cooperating was to affect one's reputation in a way that would lead to a better outcome. Thus game

end p.265

theory appears to predict both sides will also defect on play 99. This reasoning is also available to the players, and so they should also defect on play 98. Indeed, this chain of reasoning applies round by round, all the way back to the first play. Therefore there should be no cooperation at all!

What actually happened in the experiment is that one player cooperated on play 99 and the other defected. Prior to round 99 there was mostly cooperation. Subsequent experiments have found that people cooperate on most plays of an iterated Prisoners' Dilemma. Game theorists are chagrined by how ordinary people cooperate to achieve higher payoffs than is theoretically possible.

A majority of economists accept the soundness of the reasoning in the iterated Prisoners' Dilemma. They concede that in ordinary circumstances, iterating Prisoners' Dilemma leads to more cooperation. But that is only because the ordinary agents do not have the certainty enjoyed by the hypothetical prisoners. Ordinary agents are a bit unsure about whether the conditions of the game really hold. In general, the economists draw the moral that the assumption that the situation is “common knowledge” between the players is a surprisingly strong condition that is rarely satisfied in our ordinary dealings with one another. The *Alice in Wonderland* atmosphere of Prisoners' Dilemma just reflects the hidden strength of this apparently innocuous simplification.

9. Medieval Backward Inductions?

Since the iterated Prisoners' Dilemma acquired notoriety, there has been a wave of backward inductions: the chain-store paradox, the centipede, and so on (see Bicchieri, chap. 10, this volume). Instead of reviewing these spin-offs, I will consider two precursors of this influential slippery-slope argument.

Robert Louis Stevenson's 1893 short story "The Bottle Imp" features a work of the Devil: an indestructible bottle that contains an imp. The imp grants the bottle's owner nearly anything he wishes. The catch is that anyone who dies while still possessing the bottle goes to hell. If you try to throw the bottle away, it comes back. The only way to end possession is to sell the bottle at a cheaper price to someone who is well informed about the bottle's terms of ownership. In Stevenson's tale, the bottle starts out at an enormous price. It passes through famous hands: Napoleon, Captain Cook, etc. Eventually a boatswain agrees to buy the bottle for the price of two centimes, a centime being the lowest unit of currency in the world. When the seller dutifully points out that the buyer is sure to go to hell, the boatswain explains that he is going to hell anyway.

But what if the Devil had the foresight to require that the buyer not already be damned? Richard Sharvey (1983) notes that if we close such loopholes, we get a "proof" that the bottle cannot be sold for any amount of money. The bottle cannot be sold if it was bought at the lowest unit of currency, say, one centime. Thus no well-informed person will buy the bottle for two centimes. And if no one will buy at two centimes, no one will buy at three centimes. And so up to any amount of money! Yet it seems that some rational and well-informed person would buy it for a million centimes.

In the first edition of "The Bottle Imp," Stevenson erroneously credits the idea to "B. Smith." This is the garbled nickname of Richard John Smith, who was an actor in a play entitled "The Bottle Imp" (Beach 1910). The basic plot is in German folklore. The Brothers Grimm borrow it from Johann Jakob Christoffel von Grimmelshausen's *Die Landstortzerin Courasche,* published in 1756 (Starck 1911). This suggests that the idea originated as fable or a medieval legend. Since the plot was used in stories all over Europe, the insight behind Richard Sharvey's backward induction argument may be a thousand years old.

A more immediate precursor to the iterated Prisoners' Dilemma appears to have been discovered during World War II by the Swedish mathematician Lennart Ekbom (Sorensen 1988 , 253). The "prediction paradox" was in discussion at Princeton University shortly after the war (at least among Kurt Gödel and his students), and so may have reached the ears of Flood or Dresher and thereby primed the discovery of the iterated Prisoners' Dilemma. Nowadays the prediction paradox is most frequently posed as follows: A teacher announces that there will be a surprise test next week. A clever student protests that the surprise test is impossible: "Everybody knows we meet three times next week: Monday, Wednesday, and Friday. If the test were given on Friday, then on Thursday we could foresee the day of the "surprise" test. We can therefore eliminate the possibility that the test will be on Friday. This reasoning is available to us on Tuesday night. So if a test has yet to occur by then, we would know the test is on Wednesday. Only Monday remains. Therefore, a test on Monday would not be a surprise. This chain

of reasoning would be available to us on Sunday, so a Monday test would be expected. Therefore, the announced surprise test is impossible."
end p.267

The first two articles published on the prediction paradox accepted the clever student's reasoning. But since 1950, each of the subsequent hundred or so commentators has rejected the student's reasoning. They think the argument fails even if the agents are idealized in the manner envisaged by economists. The philosophers' unanimous verdict of unsoundness contrasts resoundingly with the widespread acceptance of backward induction arguments by economists.
Although there is a consensus that the clever student's reasoning is unsound, philosophers disagree about how it specifically fails. Many commentators connect the prediction paradox to G. E. Moore's problem of explaining what is odd about saying, "It is raining but I do not believe it." Although this sentence is consistent, I cannot rationally believe it. This is odd because the sentence merely describes an error on my part. As underscored by the preface paradox, I can believe that some belief or other of mine is mistaken. But I cannot believe a specific proposition is falsely believed by me. This asymmetry of believability suggests a solution to the prediction paradox. The teacher's announcement merely guarantees that the clever student will fail to have a true belief about the test date. The student can rationally believe this announcement. However, with the accumulation of test-less dates, the announcement's implications become more specific. By Thursday night, the teacher's announcement implies, "The test is on Friday but you do not believe it." Since the student cannot believe this specific attribution of an error of omission, he can no longer know that the announcement is true. Therefore, if the test is on Friday, he will be surprised by it.

10. Reflexive Beliefs

Many people who are introduced to the surprise test paradox make the following observation: If the students are persuaded by the argument, then they will not expect a test. Therefore, if the teacher gives the test, it will be a surprise. It is self-defeating to believe that the test cannot be given.
When surprise is your goal, you try to predict what your target will believe. The main basis for prediction is the principle of charity. This principle instructs us to interpret agents as rational agents. We try to make their beliefs and desires cohere. This is largely a matter of ensuring that their beliefs come out as largely true. Since we must judge truth by our own lights, this winds up as a policy of maximizing agreement. A problem arises when the goal is to disagree with those whom we are trying to interpret.
I introduce my students to this discoordination dilemma with a multiple choice question: Which answer will be chosen by the fewest classmates?
end p.268

(a)
(b)
(c)
(d)
(e)

Students think of a reason why one answer would be the least popular. But this reasoning is undermined by the likelihood that other students would replicate the reasoning.
Travelers face a less pure version of this problem when trying to avoid congestion. They try to pick unpopular times and routes but worry that "great minds think alike."
A prediction is reflexive if the probability of the prediction is affected by its conditions of dissemination. When President Bush took office in 2001, he predicted an economic downturn. People reacted in ways that helped the prediction come true.
Reflexive predictions vary in how sensitive they are to subjective reactions to them. At the logical limit, and perhaps as a degenerate case, lies one of Jean Buridan's sophisms,
(B) You do not believe this sentence.
If you do not believe the sentence, it is true and so you should believe it. Thus you cannot have a rational attitude toward the sentence. Others can. If their observation shows that you believe the sentence, then they rationally believe that the sentence is false. If their observation shows that you do not believe the sentence, then they rationally believe the sentence is true.
Buridan's sentence can be impersonalized. If a rational person believes "No rational person believes this sentence" then it is true and so merits belief. But that belief would make it false and so warrants disbelief in it.
The truth-teller analogue of this paradox is also paradoxical (Cave 2001). If you believe "You believe this statement," then your belief must be true. But can you believe it? If you believe it, then there is some proposition that you believe. The "this" in "You believe this statement" generates another this if we try to make the proposition the result of substituting the sentence "You believe this statement." Since the "this" cannot be unpacked, some conclude that "You believe this statement" is meaningless. The same objection would apply to Buridan's "You do not believe this statement."
In the literature on the liar paradox, there is a precedent for this objection and a precedent for its solution (Quine 1987 , 148). Consider
(B) "Does not yield a statement you believe when appended to its own quotation" does not yield a statement you believe when appended to its own quotation.
end p.269

Since the thirteen-word quotation is a noun rather than a demonstrative, it does not lead to the unpacking regress.
However, there still seems to no evidential basis to believe "I believe this statement." Consider a patient who realizes he is taking a placebo. He has read that even people who realize that they are taking a placebo tend to get better in virtue of their belief that they will get better. So he believes he will recover because of his belief that he will recover. Optimism in his recovery can be sustained by a preexisting belief that he will recover.

But how does he get his original belief going? That original belief seems to be without any evidential base. Although we can view others as believing *p* without the benefit of evidence, we cannot picture ourselves as believing *p* without any evidence. “I have no evidence that God exists but I believe that God exists” has the same peculiarity as Moore's sentence “It is not raining but I believe it is raining.”

11. Preference Gaps

Strangely, Jean Buridan is most famous for a paradox there is no record of him inventing. “Buridan's ass” is indifferent between two bales of hay. If a choice of *x* over *y* implies a preference for *x* over *y,* then the ass starves. Buridan's ass was a popular exhibit in commentaries on the principle of sufficient reason (which says nothing happens without a reason).
If we were able to starve asses by presenting them with equally appealing goods, then “Choice implies preference” would be empirically confirmed. Since we do not find asses transfixed in fatal indecision, there seems to be formidable empirical data against the principle of sufficient reason. However, Gottfried Leibniz, the famous champion of the principle, maintained it is *impossible* to starve an ass with two equal goods. For Leibniz denied there could be two equal goods. Leibniz appears to grant that his adversaries sometimes believe that two apples are equally good. But he does not take this belief in the equality of two goods as sufficient for indifference between the two apples. According to Leibniz, there is always at least a slight preference.
Biologists attribute less sensitivity to the ass. Arbitrary choice seems to underlie the erratic behavior of many animals (Driver and Humphries 1988). This is just what game theory predicts. Randomizing is a defense against predators who need to predict your hiding spot or anticipate your course of flight. This is especially true if you lack the brains to outwit predators.
end p.270

Some people are reluctant to admit that animals choose arbitrarily because they associate unpredictable action with a free will. This left-handed respect for free will leads to pessimism about the social sciences. How could there be science governing people without laws determining human choice?
Emile Durkheim answered that the unpredictability of individuals was compatible with the predictability of classes of people. An actuary cannot predict which individuals will commit suicide, but can predict what percentage of people will commit suicide.
The aggregate behavior of an individual is also predictable in this statistical sense. The butterfly's zigzag flight divides up into so many zigs and so many zags. The relative frequency of each would aid a ballistics expert in his attempt to shoot down the butterfly. Since physicists accept irreducibly statistical predictions when studying radioactive piles, most social scientists do not think that random choice is methodologically worrisome. After all, they themselves use random-control trials. A Martian studying human

psychologists would not be able to predict whether the randomizing psychologist was going to allocate a particular subject to the treated group or to the control group.

12. Weakness of Will

If I regard two incompatible alternatives as equally belief-worthy, then I cannot believe one over the other. If I regard them as equally desirable, I cannot desire one over the other. But if I regard two incompatible alternatives as equally choice-worthy, I can choose one over the other. Therefore, beliefs and desires do not determine all choices. If there is weakness of will, then there are graver gaps in belief-desire framework (see Mele, chap. 13, this volume). When a dieter sheepishly picks a brownie over a fruit salad, he insists that, all things considered, he prefers the fruit salad. Yet he chooses the brownie. Thus he seems to wittingly take the inferior alternative.

Anyone who accepts weakness of will accepts a significant limit on the predictive and explanatory power of beliefs and desires. Indeed, he seems to be undermining the very point of postulating beliefs and desires.

The prediction deficit could be filled with a supplementary theory that described when and how weakness of will takes place. There are many commonsense generalizations about the conditions under which people fall prey to temptation. Backsliders love company. The dieter is more apt to break his resolution if the temp tation is placed under his nose, without the notice of those who monitor whether he is on a diet, and among others who are partaking—especially when they themselves are falling to temptation. Social psychologists have elaborated such generalizations into a theory of eating. But the generalizations owe their success to a secondary effort to explain the dieter in terms of beliefs and desires—just different beliefs and desires than we first assumed. The weakness of will seems to disappear.

Although existence of weak-willed acts seems to be a truism, it is difficult to make detailed sense of it. Economists have an especially strong tendency to follow the saying "Actions speak louder than words": the dieter's choice shows that he actually prefers the brownie over the fruit salad!

13. Ill-Structured Preferences

Economists are notoriously nonjudgmental about the content of preferences. They deny a priori that the ultimate goals of Osama bin Laden could be irrational. Most economists accept David Hume's principle that "reason is and ought to be the slave of the passions." The economists share commonsense fussiness about the structure of our preferences. Almost everyone demands indifference between alternatives that are logically equivalent. People are vulnerable to framing effects in which logically equivalent alternatives are described differently. Patients and physicians are more likely to choose a treatment that is couched in terms of chances of survival rather than chances of death. These preferences

do not survive exposure. Once we realize that we prefer one alternative over a logical equivalent, we lapse into indifference.

Some of our questionable preferences about risk are more robust. Most Americans know that mile for mile, they are far more likely to be injured in a car than in an airplane. Yet the greater control afforded by a car leads them to frequently choose an eight-hour car ride over a one-hour plane trip. An economist might try to construe this as a rational choice by postulating a taste for control. But those less committed to making consumers come out rational will condemn such travelers as self-destructively muddled.

Other robust preferences about risk are more difficult to dismiss. The St. Petersburg Paradox features a fair coin that will be tossed until a head results. You will then be paid $\$2^{n-1}$ where n equals the number of tosses. So the expected return is:

$(1/2 \times \$1) + (1/4 \times \$2) + (1/8 \times \$4) ++ (1/2n \times \$2^{n-1}) +$

end p.272

Since each addend equals a half dollar, and there are infinitely many of them, the sum is infinite. Thus someone who maximized expected money should be willing to pay any amount of money for this bet. Yet few people would pay $100 for the deal.

Daniel Bernoulli (1738) denied that the principle of maximizing expected utility implies that the deal is of infinite value (even if infinite money were possible). He pointed out that doubling one's cash holdings from 1 million to 2 million does not really double its value to you. Each new dollar tends to have less influence on your welfare than the preceding dollar. Bernoulli's insight is enshrined in contemporary economics as the law of the diminishing marginal utility of money. The rate of diminution resists precise calculation, but Bernoulli inferred that it is a logarithmic function. This would preclude infinite sums.

There is more to the psychology of risk than the diminishing marginal utility of money. Maurice Allais (1979) emphasizes that people also care about the dispersion of the possible payoffs (see Dreier, chap. 9, this volume). The St. Petersburg bettor has possible future selves spread out into a mass of paupers and a mass of tycoons with hardly anything in between. This aversion to wide ranging outcomes is manifested most strikingly in our taste for certainty. Daniel Kahneman and Amos Tversky (1979) firmed up the empirical basis for the “Allais paradox” with a number of experiments in Israel. At the time of their study, the median net monthly income for a family was 3,000 Israeli pounds. Israelis were asked for their preferences between the following options:

Option A: 4000 with a probability of .8
Option B: 3000 with certainty
Option C: 4000 with a probability of .2
Option D: 3000 with a probability of .25

The majority of Israelis preferred B over A even though A has a higher expected return of 3,200. This cannot be explained by the diminishing marginal utility of money, because many of the same people prefer C (expected return 800) over D (expected return 750).

People also dislike “risky” processes. Daniel Ellsberg (1961) has us consider two lotteries. In the first lottery, you get a $100 prize if you choose a red marble from an urn that you know to be composed of 50 red marbles and 50 black marbles. In the second lottery, the urn contains red marbles and black marbles in an unknown ratio. You get $100 if you pick a red marble. Standard decision theory says you should be indifferent between the two lotteries because they yield the same expected return. Yet a large percentage of people prefer the first lottery. Many continue to maintain the preference for the better known urn even after the equal expected return is made salient.
There are important anomalies of preference that are independent of risk. Breakdowns of the transitivity of preference are the most discussed (see Dreier,
end p.273

chap. 9, and Joyce, chap. 8, this volume). If a shopper prefers car A over car B and car B over car C, then he *should* prefer car A over car C. If the shopper instead prefers car C over car A, then the car salesman will be struck by the “inconsistency.” He might try to prove the irrationality of the shopper by converting him into a money pump. If the shopper prefers C over A, then he should be willing to trade A plus a dollar to get C. Since he prefers B over C, the shopper should trade C plus a dollar to get B. And since he prefers A over B, the shopper should trade B plus a dollar to get A. Although each of the three trades is fair (the shopper always swaps for a more preferred basket of goods), the shopper winds up in the same position (owning car A) minus three dollars. The absurdity can be magnified by repeating the cycle over and over; all the shopper's money could be drained away by the car salesman.

14. Social Choice

Transitivity also breaks down for collective preferences. The marquis de Condorcet has us imagine that a group of three individuals is governed by majority rule. Their preferences are individually transitive:

Voter 1: A, B, C
Voter 2: B, C, A
Voter 3: C, A, B

Yet the group's preference is not transitive. By majority rule, the group prefers A over B (because voters 1 and 3 vote for A over B), B over C (because voters 1 and 2 vote for B over C), and C over A (because voters 2 and 3 vote for C over A). Given this cyclic preference structure, the group decision will depend on the order in which the issues are voted upon. If we start by asking whether A or B should be eliminated, the group will eliminate B. If we then continue the deliberation by asking whether A or C should be eliminated, then A will be eliminated and the group will ultimately decide in favor of C.

But had the first stage of the vote begun with the question of whether B or C should be eliminated, then C would have been eliminated and A would have ultimately prevailed. Democrats offered increasingly complex repairs to this vulnerability to sophisticated voting. Kenneth Arrow (1951) ended the quest for a perfect “social aggregation device” by proving that all democratic voting schemes must form some cyclical preferences. Perfect democracy is impossible—and not merely because it must be shaped with imperfect human clay.
end p.274

Condorcet's insight about cyclical majorities applies to individual people insofar as they are pictured as collectives. Consider self-deception (see Mele, chap. 13, this volume). If Cicero deceives Tully into believing *p,* then Tully must believe. Yet it must also be the case that Cicero does not believe. After all, if Cicero believed *p,* then he would be trying to lead Tully to a truth, not a falsehood. These two requirements preclude the possibility that Cicero is identical to Tully. For if Cicero is Tully, then Cicero does not believe *p* (to qualify as a deceiver) and does believe *p* (to qualify as the victim of the deceiver). A common solution is to say that part of the self-deceiver believes *p* and part of him does not. These parts are homunculi, “little men” who are capable of deceiving each other. This homuncular model is the norm in explanations of illusions. If we are collectives of homunculi, then it is probable that some of our preferences are formed by majority rule. Hence, even individuals should be vulnerable to Condorcet's paradox. Indeed, the contemporary revival of faculty psychology suggests that any paradox of collective rationality should arise in miniature at the level of individuals.

NOTE

I thank Andreas Teaber for his Stevenson scholarship.

part ii RATIONALITY IN SPECIFIC DOMAINS
end p.277

end p.278

chapter 15 RATIONALITY AND PSYCHOLOGY
Richard. Samuels
Stephen. Stich

Since the early 1970s, psychologists have devoted a great deal of attention to human reasoning and decision making, and to the psychological processes that underlie them. While some of this attention was motivated by the intrinsic interest and importance of

these processes, much of it was provoked by a series of experimental findings that, in the view of many, had “bleak implications” for human rationality (Nisbett and Borgida 1975). In this essay we'll begin, in section 1, by presenting a brief sketch of some of these disturbing findings, most of which were reported by psychologists in what has become known as the “heuristics and biases” tradition. In section 2, we'll set out three increasingly pessimistic interpretations of the findings that have been suggested by a number of authors. There have been many challenges to these pessimistic interpretations. One of the most interesting and influential challenges was launched, in the early 1990s, by a group of researchers working in the then newly emerging interdisciplinary field of evolutionary psychology. This challenge, and the experimental findings that support it, will be our focus in section 3. On the basis of these new findings, evolutionary psychologists have suggested a variety of much more optimistic views about the rationality of ordinary people. In section 4, we'll sketch three of these views. We are inclined to think that the right reaction to the entire body of findings on human reasoning is significantly less pessimistic than the most dire interpretation
end p.279

suggested by writers in the heuristics and biases tradition, but rather more pessimistic than suggested by the Panglossian pronouncements of some evolutionary psychologists. In section 5, we'll defend this “middle way” and sketch a family of “dual processing” theories of reasoning that, we'll argue, offer some support for the moderate interpretation we advocate.

1. Some Unsettling Studies of Human Reasoning

In 1966, Peter Wason published a highly influential study of a cluster of reasoning problems that became known as the *selection task*. By 1993, the selection task had become “the most intensively researched single problem in the history of the psychology of reasoning.” (Evans, Newstead, and Byrne 1993 , 99) Figure 15.1 illustrates a typical example of a selection task problem. What Wason and numerous other investigators have found is that subjects typically perform very poorly on questions like this. Most subjects respond correctly that the E card must be turned over, but many also insist that the 5 card must be turned over, though the 5 card could not falsify the claim no matter what is on the other side. Also, a majority of subjects maintain that the 4 card need *not* be turned over, though without turning it over there is no way of knowing whether it has a vowel on the other side. Subjects do not do poorly on *all* selection task problems, however. A wide range of variations on the basic pattern has been tried, and on some versions of the problem a much larger percentage of subjects answers correctly. These results form a bewildering pattern, since there is no obvious feature or cluster of features that separates versions on which subjects do well from those on which they do poorly.
Much of the experimental literature on theoretical reasoning has focused on tasks that concern *probabilistic* judgment. Among the best-known experiments of this kind are

those that involve so-called *conjunction problems*. In one quite famous experiment, Tversky and Kahneman (1982 , p. 92) presented subjects with the following task.
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations.
end p.280

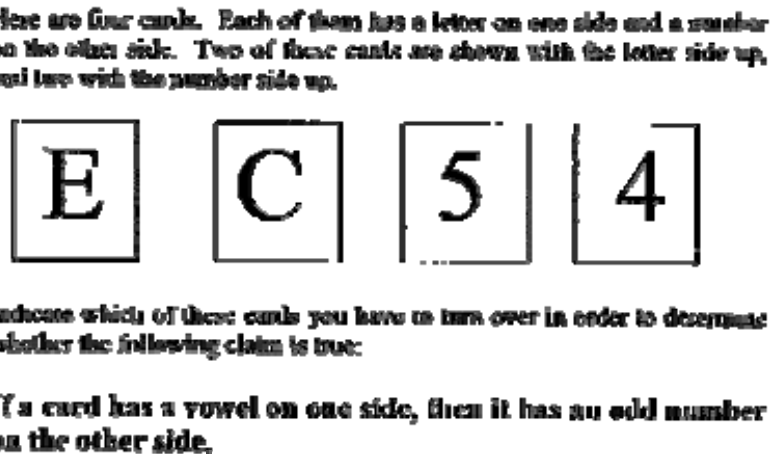

Figure 15.1

Please rank the following statements by their probability, using 1 for the most probable and 8 for the least probable.
(a) Linda is a teacher in elementary school.
(b) Linda works in a bookstore and takes Yoga classes.
(c) Linda is active in the feminist movement.
(d) Linda is a psychiatric social worker.
(e) Linda is a member of the League of Women Voters.
(f) Linda is a bank teller.
(g) Linda is an insurance sales person.
(h) Linda is a bank teller and is active in the feminist movement.

In a group of naive subjects with no background in probability and statistics, 89 percent judged that (h) was more probable than (f) despite the obvious fact that one cannot be a *feminist* bank teller unless one is a *bank teller*. When the same question was presented to statistically sophisticated subjects—graduate students in the decision science program of the Stanford Business School—85 percent gave the same answer! Results of this sort, in which subjects judge that a compound event or state of affairs is more probable than one of the components of the compound, have been found repeatedly since Kahneman and Tversky's pioneering studies, and they are remarkably robust. This pattern of reasoning has been labeled the *conjunction fallacy*.
Another well-known cluster of studies examines the way in which people use base-rate information in making probabilistic judgments. According to the familiar Bayesian

account, the probability of a hypothesis on a given body of evidence depends, in part, on the prior probability of the hypothesis. However, in a series of elegant experiments, Kahneman and Tversky (1973) showed that subjects often seriously undervalue the importance of prior probabilities. One of these experiments presented half of the subjects with the following "cover story."

A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100. (p. 54)

The other half of the subjects was presented with the same text, except the "base-rates" were reversed. These subjects were told that the personality tests had been administered to 70 engineers and 30 lawyers. Some of the descriptions that were provided were designed to be compatible with the subjects' stereotypes of engineers, though not with their stereotypes of lawyers. Others were designed to fit the lawyer stereotype, but not the engineer stereotype. And one was intended to be quite neutral, giving subjects no information at all that would be of use in making their decision. Here are two examples, the first intended to sound like an engineer, the second intended to sound neutral:

Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies, which include home carpentry, sailing, and mathematical puzzles.

Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues. (p. 54)

As expected, subjects in both groups thought that the probability that Jack is an engineer is quite high. Moreover, in what seems to be a clear violation of Bayesian principles, the difference in cover stories between the two groups of subjects had almost no effect at all. The neglect of base-rate information was even more striking in the case of Dick. That description was constructed to be totally uninformative with regard to Dick's profession. Thus, the *only* useful information that subjects had was the base-rate information provided in the cover story. But that information was entirely ignored. The median probability estimate in both groups of subjects was 50 percent.

Before leaving the topic of base-rate neglect, we want to offer one further example illustrating the way in which the phenomenon might well have serious practical consequences. Here is a problem that Casscells, Schoenberger, and Gray

boys (1978 , p. 999) presented to a group of faculty, staff, and fourth-year students at Harvard Medical School.

If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs? %

Under the most plausible interpretation of the problem, the correct Bayesian answer is 2 percent. But only 18 percent of the Harvard audience gave an answer close to 2 percent. Forty-five percent of this distinguished group completely ignored the base-rate information and said that the answer was 95 percent.
One of the most extensively investigated and most worrisome cluster of phenomena explored by psychologists interested in reasoning and judgment involves the degree of confidence that people have in their responses to factual questions—questions like:
In each of the following pairs, which city has more inhabitants?

(a) Las Vegas (b) Miami
(a) Sydney (b) Melbourne
(a) Hyderabad (b) Islamabad
(a) Bonn (b) Heidelberg

In each of the following pairs, which historical event happened first?

(a) Signing of the Magna Carta (b) Birth of Mohammed
(a) Death of Napoleon (b) Louisiana Purchase
(a) Lincoln's assassination (b) Birth of Queen Victoria

After each answer subjects are also asked:
How confident are you that your answer is correct?
50% 60% 70% 80% 90% 100%
In an experiment using relatively hard questions it is typical to find that for the cases in which subjects say they are 100 percent confident, only about 80 percent of their answers are correct; for cases in which they say that they are 90 percent confident, only about 70 percent of their answers are correct; and for cases in which they say that they are 80 percent confident, only about 60 percent of their answers are correct. This tendency toward overconfidence seems to be very robust. Warning subjects that people are often overconfident has no significant effect, nor does offering them money (or bottles of Champagne) as a reward for accuracy.
end p.283

Moreover, the phenomenon has been demonstrated in a wide variety of subject populations including undergraduates, graduate students, physicians, and even CIA analysts. (For a survey of the literature, see Lichtenstein, Fischoff, and Phillips 1982 .)
The studies we've reviewed so far have focused on subjects' normatively problematic performance on belief formation and judgmental tasks. But there is also a large experimental literature that seems to indicate that human decision making processes are normatively problematic. Since space is limited, we'll recount only one example, albeit a

particularly disturbing one. Tversky and Kahneman (1981) presented a group of subjects with the following problem:

Imagine that the U.S. is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs is as follows:

If Program A is adopted, 200 people will be saved.

If Program B is adopted, there is a 1/3 probability that 600 people will be saved, and a 2/3 probability that no people will be saved.

A second group of subjects was given an identical problem, except that the programs were described as follows:

If Program C is adopted, 400 people will die.

If Program D is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 people will die.

On the first version of the problem a substantial majority of the subjects chose program A. But on the second version most chose program D, despite the fact that the outcome described in A is identical to the one described in C. The disconcerting implication of this study is that the decisions we make are strongly influenced by the manner in which the options are described or *framed*.

2. What Do These Results Show? Three Pessimistic Views

What do these results and the many similar results to be found in the experimental literature [1] tell us about the rationality of ordinary people's reasoning and decision

end p.284

making and about the mental mechanisms that underlie those processes? In this section we'll distinguish three answers to this question that have been suggested in the literature—answers that get increasingly pessimistic (and, as we will argue in section 5, increasingly implausible).

Before attempting to the answer the question, we must, of course, adopt some normative standard or metric for assessing the rationality of inferences and decisions. Though researchers in this area rarely offer an explicit and general normative theory of rationality, we think that most authors tacitly adopt some version of what Edward Stein has called 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 so forth. If the standard picture of reasoning is right, principles of reasoning that are based on such rules are normative principles of reasoning, namely they are the principles we ought to reason in accordance with” (Stein 1996 , 4).

Thus the Standard Picture maintains that the appropriate criteria against which to evaluate human reasoning are the rules derived from formal theories such as classical logic, probability theory, and decision theory. [2] So, for example, one might derive something like the following principle of reasoning from the conjunction rule of probability theory:

Conjunction Principle: One ought not assign a lower degree of probability to the occurrence of event A than one does to the occurrence of A and some (distinct) event B. (Stein 1996 , 6)

Given principles of this kind, one can evaluate the judgments and decisions of human subjects and the mechanisms that produce them. To the extent that a person's judgments and decisions accord with the principles of the Standard Picture, they are rational, and to the extent that they violate such principles, they fail to be rational. Similarly, to the extent that a reasoning or decision making mechanism produces judgments that accord with the principles of the Standard Picture, the mechanism is rational and to the extent that it fails to do so, it is not rational.

If we adopt the Standard Picture, then one quite plausible conclusion to draw from the experimental findings reported in the heuristics and biases literature is that

(1) People's intuitive judgments on a large number of reasoning and decision making problems regularly deviate from appropriate norms of rationality.

To understand a second claim that has been made on the basis of the experimental findings, we need to recount how researchers in the heuristics and biases tradition explain people's performance on many reasoning problems. The basic

end p.285

explanatory strategy is to posit the existence of reasoning "heuristics"—rules of thumb that people employ when reasoning. In the case of the conjunction fallacy and base rate neglect, for example, Kahneman and Tversky propose that people often rely on what they call *the representativeness heuristic*.

Given specific evidence (e.g. a personality sketch), the outcomes under consideration (e.g. occupations or levels of achievement) can be ordered by the degree to which they are representative of that evidence. The thesis of this paper is that people predict by representativeness, that is, they select or order outcomes by the degree to which the outcomes represent the essential features of the evidence. In many situations, representative outcomes are indeed more likely than others. However, this is not always the case, because there are factors (e.g. prior probabilities of outcomes and the reliability of evidence) which effect the likelihood of outcomes but not their representativeness. Because these factors are ignored, intuitive predictions violate statistical rules of prediction in systematic and fundamental ways. (Kahneman and Tversky 1973 , 48)

If explanations of this sort are correct, then we can conclude that:

(2) Many of the instances in which our judgments and decisions deviate from appropriate norms of rationality are explained by the fact that, in making these judgments and decisions, people rely on heuristics like representativeness "which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors" (Kahneman and Tversky 1973 , 48).

Though (1) and (2) have been challenged in a number of ways, they are both relatively modest reactions to the experimental findings. [3] However, many writers have suggested a much stronger and more disturbing conclusion, which maintains that people use these heuristics *because they have no better tools available for dealing with many reasoning and decision making problems*. According to Slovic, Fischhoff and Lichtenstein, for example, "It appears that people lack the correct programs for many important

judgmental tasks.We have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty" (1976 , 174). What they appear to be suggesting is that:

(3) The *only* cognitive tools that are available to untutored people are normatively problematic heuristics such as representativeness.

This pessimistic conclusion seems to be endorsed in passages like the following in which Kahneman and Tversky, the founders of the heuristics and biases tradition, maintain that people use representativeness and other normatively defective heuristics not just in some or many cases but in *all* cases—including those cases in which they get the right answer: "In making predictions and judgments under uncertainty, people do not appear to follow the calculus of chance or the statistical theory of prediction. Instead, they rely on a limited number of heuristics which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors" (Kahneman and Tversky, 1973 , 48). In light of passages like this, it is hardly surprising that both friends and foes of the heuristics and biases tradition suppose that it is committed to the claim that, as Gerd Gigerenzer has put it, "the untutored mind is running on shoddy software, that is, on programs that work *only* with a handful of heuristics" (1991b , 235). After describing the feminist bank teller experiment, the eminent biologist Stephen J. Gould, who is an admirer of work in the heuristics and biases tradition, asks: "Why do we consistently make this simple logical error?" His answer is: "Tversky and Kahneman argue, correctly I think, that our minds are not built (for whatever reason) to work by the rules of probability" (1992 , 469). [4] And, making the point a bit more bluntly, Lola Lopes, a psychologist who has been a trenchant critic of the heuristics and biases tradition, has suggested that researchers in that tradition think "that people are pretty stupid" (Lopes, quoted in Bower 1996).

3. Evolutionary Psychologists' Critique of the "Shoddy Software" Hypothesis

The hypothesis that the only cognitive tools available to most human reasoners and decision makers are "shoddy software" like the representativeness heuristic has been the main target of an important critique of the heuristics and biases tradition mounted by evolutionary psychologists. In this section we'll give an overview of that critique. Though the interdisciplinary field of evolutionary psychology is too new to have developed any precise and widely agreed upon body of doctrine, there are three basic theses that are clearly central. The first is that the mind contains a large number of special purpose systems—often called "modules" or "mental organs." These modules are invariably conceived of as a type of computational mechanism, namely, computational devices that are specialized or domain specific. Many evolutionary psychologists also urge that modules are both innate and present in all normal members of the species. While this characterization of modules raises lots of interesting issues—about which we have had a fair amount to say elsewhere (Samuels 1998 ; Samuels, Stich, and Tremoulet 1999 ; Samuels 2000)—in this essay we propose to put them to one side. The second central thesis of evolutionary psychology is that, contrary to what has been argued by Fodor (1983) and others, the

modular structure of the mind is not restricted to input systems (those responsible for perception and language processing) and output systems (those responsible for producing actions). According to evolutionary psychologists, modules also subserve many so-called central mental capacities such as reasoning, desire formation and decision making. [5] The third thesis is that mental modules are *adaptations*—they were, as Tooby and Cosmides have put it, "invented by natural selection during the species' evolutionary history to produce adaptive ends in the species' natural environment" (Tooby and Cosmides 1995 , xiii).

The Frequentist Hypothesis

If much of central cognition is indeed subserved by cognitive modules that were designed to deal with the adaptive problems posed by the environment in which our primate forebears lived, then we should expect that the modules responsible for reasoning would do their best job when information is provided in a format similar to the format in which information was available in the ancestral environment. And, as Gigerenzer has argued, though there was a great deal of useful probabilistic information available in that environment, this information would have been represented "as frequencies of events, sequentially encoded as experienced—for example, *3 out of 20* as opposed to 15% or p = 0.15" (1994 , 142). Cosmides and Tooby make much the same point: "Our hominid ancestors were immersed in a rich flow of observable frequencies that could be used to improve decision-making, given procedures that could take advantage of them. So if we have adaptations for inductive reasoning, they should take frequency information as input" (1996 , 15–16). On the basis of such evolutionary considerations, Gigerenzer, Cosmides, and Tooby have proposed and defended a psychological hypothesis that they refer to as the *Frequentist Hypothesis:* "Some of our inductive reasoning mechanisms do embody aspects of a calculus of probability, but they are designed to take frequency information as input and produce frequencies as output" (Cosmides and Tooby 1996 , 3). This speculation led Cosmides and Tooby to pursue an intriguing series of experiments in which the "Harvard Medical School problem" used by Casscells et al. was systematically transformed into a problem in which both the input and the response required were formulated in terms of frequencies. Here is one example from their study in which frequency information is made particularly salient:

1 out of every 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely

healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease.
Imagine that we have assembled a random sample of 1000 Americans. They were selected by lottery. Those who conducted the lottery had no information about the health status of any of these people.
Given the information above:
on average,
How many people who test positive for the disease will *actually* have the disease? out of .
(p. 24)
In sharp contrast to Casscells et al.'s original experiment, in which only eighteen percent of subjects gave the correct Bayesian response, this problem elicited the correct Bayesian answer from 76 percent of Cosmides and Tooby's subjects.
A series of further experiments systematically explored the differences between the problem used by Casscells et al. and the problems on which subjects perform well, in an effort to determine which factors had the largest effect. Although a number of different factors affect performance, two predominate: "Asking for the answer as a frequency produces the largest effect, followed closely by presenting the problem information as frequencies" (Cosmides and Tooby 1996 , 58). The most important conclusion that Cosmides and Tooby want to draw from these experiments is that "frequentist representations activate mechanisms that produce bayesian reasoning, and that this is what accounts for the very high level of bayesian performance elicited by the pure frequentist problems that we tested" (59).
As further support for this conclusion, Cosmides and Tooby cite several striking results reported by other investigators. In one study, Fiedler 1988 , following up on some intriguing findings in Tversky and Kahneman 1983 , showed that the percentage of subjects who commit the conjunction fallacy can be radically reduced if the problem is cast in frequentist terms. In the "feminist bank teller" example, Fiedler contrasted the wording reported in section 1 with a problem that read as follows:
Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations.
There are 100 people who fit the description above. How many of them are:
bank tellers?
bank tellers and active in the feminist movement?
(p. 125)
end p.289

In Fiedler's replication using the original formulation of the problem, 91 percent of subjects judged the feminist bank teller option to be more probable than the bank teller option. However, in the frequentist version, only 22 percent of subjects judged that there would be more feminist bank tellers than bank tellers. In yet another experiment, Hertwig and Gigerenzer (1994 ; reported in Gigerenzer 1994) told subjects that there were two hundred women fitting the "Linda" description, and asked them to estimate the number who were bank tellers, feminist bank tellers, and feminists. Only 13 percent committed the conjunction fallacy.

Studies on overconfidence have also been marshaled in support of the frequentist hypothesis. In one of these Gigerenzer, Hoffrage, and Kleinbölting (1991) reported that the sort of overconfidence described above can be made to “disappear” by having subjects answer questions formulated in terms of frequencies. Gigerenzer and his colleagues gave subjects lists of 50 questions similar to those described in section 1, except that in addition to being asked to rate their confidence after each response (which, in effect, asks them to judge the probability of that single event), subjects were, at the end, also asked a question about the frequency of correct responses: “How many of these 50 questions do you think you got right?” In two experiments, the average overconfidence was about 15 percent when single-event confidences were compared with actual relative frequencies of correct answers, replicating the sorts of findings we sketched in section 1. However, comparing the subjects' “estimated frequencies with actual frequencies of correct answers made “overconfidence” *disappear*.Estimated frequencies were practically identical with actual frequencies, with even a small tendency towards underestimation. The “cognitive illusion” was gone” (Gigerenzer 1991a , 89).

The Cheater Detection Hypothesis

In section 1 we reproduced one version of Wason's selection task on which most subjects perform very poorly, and we noted that, while subjects do equally poorly on many other versions of the selection task, there are some versions on which performance improves dramatically. Figure 15.2 is an example from Griggs and Cox 1982 . From a logical point of view, this problem would appear to be quite similar to the problem in section 1, but the *content* of the problems clearly has a major effect on how well people perform. About 75 percent of college student subjects get the right answer on this version of the selection task, while only 25 percent get the right answer on the other version. Though there have been dozens of studies exploring this “content effect” in the selection task, the results have been and continue to be rather puzzling since there is no obvious property or set of properties shared by those versions of the task on which people perform well.
end p.290

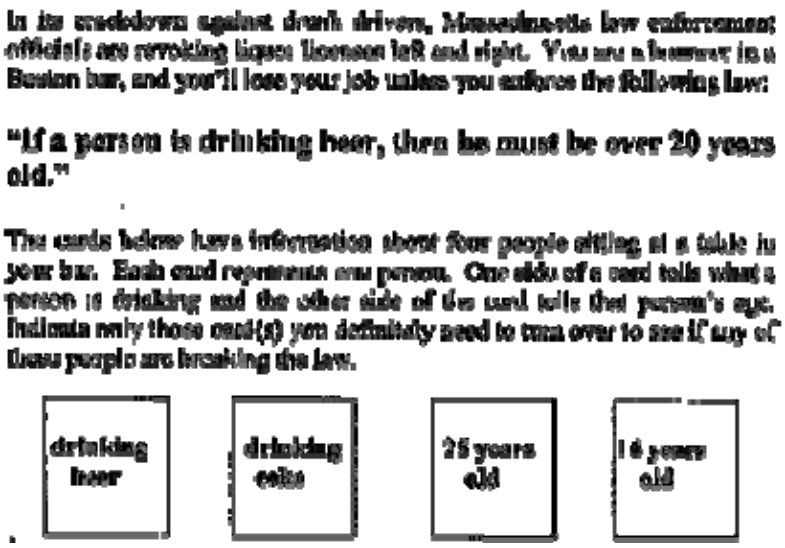

Figure 15.2

However, in several widely discussed papers, Cosmides and Tooby have argued that an evolutionary analysis enables us to see a surprising pattern in these otherwise bewildering results (Cosmides 1989 , Cosmides and Tooby, 1992).

The starting point of their evolutionary analysis is the observation that in the environment in which our ancestors evolved (and in the modern world as well) it is often the case that unrelated individuals can engage in "non-zero-sum" exchanges, in which the benefits to the recipient (measured in terms of reproductive fitness) are significantly greater than the costs to the donor. In a hunter-gatherer society, for example, it will sometimes happen that one hunter has been lucky on a particular day and has an abundance of food, while another hunter has been unlucky and is near starvation. If the successful hunter gives some of his meat to the unsuccessful hunter rather than gorging on it himself, this may have a small negative effect on the donor's fitness, since the extra bit of body fat that he might add could prove useful in the future, but the benefit to the recipient will be much greater. Still, there is *some* cost to the donor; he would be slightly better off if he didn't help unrelated individuals. Despite this, it is clear that people sometimes do help non-kin, and there is evidence to suggest that nonhuman primates (and even vampire bats!) do so as well. On first blush, this sort of "altruism" seems to pose an evolutionary puzzle, since if a gene that made an organism *less* likely to help unrelated individuals appeared in a population, those with the gene would be slightly *more* fit, and thus the gene would gradually spread through the population.

A solution to this puzzle was proposed by Robert Trivers (1971), who noted that while one-way altruism might be a bad idea from an evolutionary point of view, *reciprocal altruism* is quite a different matter. If a pair of hunters (be they humans or bats) can each count on the other to help when one has an abundance of food and the other has none, then they may both be better off in the long run. Thus organisms with a gene or a suite of genes that inclines them to engage in reciprocal exchanges with non-kin (or "social exchanges" as they are sometimes called) would be more fit than members of the same species without those genes. But of course, reciprocal exchange arrangements are

vulnerable to cheating. In the business of maximizing fitness, individuals will do best if they are regularly offered and accept help when they need it but never reciprocate when others need help. This suggests that if stable social exchange arrangements are to exist, the organisms involved must have cognitive mechanisms that enable them to detect cheaters and to avoid helping them in the future. And since humans apparently are capable of entering into stable social exchange relations, this evolutionary analysis led Cosmides and Tooby to hypothesize that we have one or more mental modules whose job it is to recognize reciprocal exchange arrangements and to detect cheaters who accept the benefits in such arrangements but do not pay the costs. In short, the evolutionary analysis leads Cosmides and Tooby to hypothesize the existence of one or more cheater detection modules. We call this the *cheater detection hypothesis*.

If this is right, then we should be able to find some evidence for the existence of these modules in the thinking of contemporary humans. It is here that the selection task enters the picture. For according to Cosmides and Tooby, some versions of the selection task engage the mental modules that were designed to detect cheaters in social exchange situations. And since these mental modules can be expected to do their job efficiently and accurately, people do well on those versions of the selection task. Other versions of the task do not trigger the social exchange and cheater detection modules. Since we have no mental modules that were designed to deal with these problems, people find them much harder, and their performance is much worse. The bouncer-in-the-Boston-bar problem is an example of a selection task that triggers the cheater detection mechanism. The problem involving vowels and odd numbers presented in section 1 is an example of a selection task that does not trigger the cheater detection module.

In support of their theory, Cosmides and Tooby assemble an impressive body of evidence. To begin, they note that the cheater detection hypothesis claims that social exchanges, or "social contracts," will trigger good performance on selection tasks, and this enables us to see a clear pattern in the otherwise confusing experimental literature that had grown up before their hypothesis was formulated.

When we began this research in 1983, the literature on the Wason selection task was full of reports of a wide variety of content effects, and there was no satisfying theory or empirical generalization that could account for these effects. When we categorized these content effects according to whether they conformed to social contracts, a striking pattern emerged. Robust and replicable content effects were found only for rules that related terms that are recogniza

end p.292

ble as benefits and cost/requirements in the format of a standard social contract.No thematic rule that was not a social contract had ever produced a content effect that was both robust and replicable.All told, for non–social contract thematic problems, 3 experiments had produced a substantial content effect, 2 had produced a weak content effect, and 14 had produced no content effect at all. The few effects that were found did not replicate. In contrast, 16 out of 16 experiments that fit the criteria for standard social contractselicited substantial content effects. (Cosmides and Tooby, 1992 , 183)

Since the formulation of the cheater detection hypothesis, a number of additional experiments have been designed to test the hypothesis and rule out alternatives. Among

the most persuasive of these are a series of experiments by Gigerenzer and Hug (1992). In one set of experiments, these authors set out to show that, contrary to an earlier proposal by Cosmides and Tooby, *merely* perceiving a rule as a social contract was not enough to engage the cognitive mechanism that leads to good performance in the selection task, and that cueing for the possibility of *cheating* was required. To do this they created two quite different context stories for social contract rules. One of the stories required subjects to attend to the possibility of cheating, while in the other story cheating was not relevant. Among the social contract rules they used was the following, which, they note, is widely known among hikers in the Alps:

(i) If someone stays overnight in the cabin, then that person must bring along a bundle of wood from the valley.

The first context story, which the investigators call the "cheating version," explained that there is a cabin at high altitude in the Swiss Alps, which serves hikers as an overnight shelter. Since it is cold and firewood is not otherwise available at that altitude, the rule is that each hiker who stays overnight has to carry along his/her own share of wood. There are rumors that the rule is not always followed. The subjects were cued into the perspective of a guard who checks whether any one of four hikers has violated the rule. The four hikers were represented by four cards that read "stays overnight in the cabin," "carried no wood," "carried wood," and "does not stay overnight in the cabin."

The other context story, the "no cheating version,"

cued subjects into the perspective of a member of the German Alpine Association who visits the Swiss cabin and tries to discover how the local Swiss Alpine Club runs this cabin. He observes people bringing wood to the cabin, and a friend suggests the familiar overnight rule as an explanation. The context story also mentions an alternative explanation: rather than the hikers, the members of the Swiss Alpine Club, who do not stay overnight, might carry the wood.

end p.293

The task of the subject was to check four persons (the same four cards) in order to find out whether anyone had violated the overnight rule suggested by the friend. (Gigerenzer and Hug 1992 , 142–43)

The cheater detection hypothesis predicts that subjects will do better on the cheating version than on the no cheating version, and that prediction was confirmed. In the cheating version, 89 percent of the subjects got the right answer, while in the no cheating version, only 53 percent responded correctly.

In another set of experiments, Gigerenzer and Hug showed that when social contract rules make cheating on both sides possible, cueing subjects into the perspective of one party or the other can have a dramatic effect on performance in selection task problems. One of the rules they used that allows the possibility of bilateral cheating was:

(ii.) If an employee works on the weekend, then that person gets a day off during the week.

Here again, two different context stories were constructed, one of which was designed to get subjects to take the perspective of the employee, while the other was designed to get subjects to take the perspective of the employer.

The employee version stated that working on the weekend is a benefit for the employer, because the firm can make use of its machines and be more flexible. Working on the weekend, on the other hand is a cost for the employee. The context story was about an employee who had never worked on the weekend before, but who is considering working on Saturdays from time to time, since having a day off during the week is a benefit that outweighs the costs of working on Saturday. There are rumors that the rule has been violated before. The subject's task was to check information about four colleagues to see whether the rule has been violated. The four cards read: "worked on the weekend," "did not get a day off," "did not work on the weekend," "did get a day off."
In the employer version, the same rationale was given. The subject was cued into the perspective of the employer, who suspects that the rule has been violated before. The subjects' task was the same as in the other perspective [viz. to check information about four employees to see whether the rule has been violated]. (Gigerenzer and Hug 1992 , 154)
In these experiments, about 75 percent of the subjects cued to the employee's perspective chose the first two cards ("worked on the weekend" and "did not get a day off") while less than 5 percent chose the other two cards. The results for subjects cued to the employer's perspective were radically different. Over 60 percent of subjects selected the last two cards ("did not work on the weekend" and "did get a day off") while less than 10 percent selected the first two. These experiments, along with a number of others reviewed in Cosmides and Tooby 1992 ,
end p.294

are all compatible with the hypothesis that we have one or more Darwinian modules designed to deal with social exchanges and detect cheaters. [6]

4. What Do These Results Show? Three Optimistic Answers

What do *these* results tell us about the rationality of ordinary people's reasoning and decision making? Evolutionary psychologists have urged that their findings support the truth of three increasingly optimistic claims. First, they maintain, the data suggest that:
(4) There are many reasoning and decision-making problems on which people's intuitive judgments *do not* deviate from appropriate norms of rationality.
Second, they have argued that:
(5) Many of the instances in which our judgments and decisions accord with appropriate norms of rationality are to be explained by the fact that, in making these judgments, we rely on mental modules that were designed by natural selection to do a good job at nondemonstrative reasoning when provided with the sort of input that was common in the environment in which our hominid ancestors evolved.
Finally, evolutionary psychologists have also on occasion issued exuberantly Panglossian proclamations suggesting that

(6) Most or all of our reasoning and decision making is subserved by normatively unproblematic "elegant machines" designed by natural selection, and thus any concerns about systematic irrationality are unfounded.

This optimistic view is suggested in numerous places in the evolutionary psychology literature. For example, the paper in which Cosmides and Tooby reported their data on the Harvard Medical School problem appeared with the title "Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty." Five years earlier, while Cosmides and Tooby's research was still in progress, Gigerenzer reported some of their early

end p.295

findings in a paper with the provocative title "How to make cognitive illusions disappear: Beyond `heuristics and biases.'" The clear suggestion, in both of these titles, is that the findings they report pose a head-on challenge to (1), the weakest of the pessimistic conclusions that have been drawn from research in the heuristics and biases tradition. Nor were these suggestions restricted to titles. In paper after paper, Gigerenzer has said things like: "We need not necessarily worry about human rationality" (1998 , 280); "More optimism is in order" (1991b , 245); "Keep distinct meanings of probability straight, and much can be done—cognitive illusions disappear" (ibid); and he has maintained that his view "supports intuition as basically rational" (1991b, 242). In light of comments like these, many observers have concluded that the view of the mind and of human rationality proposed by evolutionary psychologists is fundamentally at odds with the view offered by proponents of the heuristics and biases program. [7]

5. Who's Right? The "Middle Way" and Dual Processing Theories

Though this is not the place to defend our view in detail, we are inclined to think that the correct conclusions to draw about the rationality of ordinary folk from the large and growing body of experimental findings about reasoning and decision making are not nearly so optimistic as (6) nor nearly so pessimistic as (3). [8] To begin, note that (1), which claims that people's intuitive judgments on many reasoning and decision making problems deviate from appropriate norms of rationality, and (4), which claims that there are many reasoning and decision making tasks on which people's intuitive judgments *do not* deviate from appropriate norms of rationality, are entirely compatible. Moreover, we believe that the evidence reviewed in sections 1 and 3, along with many other studies that might have been discussed, make an overwhelmingly plausible case that both (1) and (4) are true. People do make serious and systematic errors on many reasoning tasks, but they also perform quite well on many others. The heuristics and biases tradition has focused on the former cases, while evolutionary psychologists have focused on the latter.

The pessimistic (2) and the optimistic (5) are both explanatory hypotheses that make claims about the sorts of psychological mechanisms and processes that underlie these two sorts of cases. And here again, there is no inconsistency—both (2) and (5) might well be true. We think that each of these explanatory hypotheses may indeed turn out to be true,

though each has serious competitors. There isstill much to learn about the mechanisms and processes subserving reasoning, and until more is known we think it would be premature to either accept or reject (2) and (5).

Though (1), (2), (4), and (5) might all be true, it can't be the case that both (3) and (6) are true, since the former insists that the only cognitive tools most people have available are normatively problematic heuristics that will lead to systematic errors, while the latter maintains that most reasoning is subserved by normatively unproblematic "elegant machines," and thus that we need not worry about human rationality. But while (3) and (6) can't both be true, they can both be false, and we believe they are. There is nothing in the heuristics and biases literature that supports the claim that problematic heuristics are the *only* reasoning resources people can draw on, nor does this literature provide us with any reason to think that normatively unproblematic mechanisms like those posited by evolutionary psychologists do not exist. On the other side, evolutionary psychologists certainly have offered no reason to think that *all* reasoning is subserved by normatively unproblematic modules. Indeed, if it is granted, as we think it should be, that people typically do poorly on a large and important class of reasoning problems, then it is clear that (6) is indefensible.

We believe that the "middle way" we've been urging between the pessimism suggested by the heuristics and biases tradition and the optimism proclaimed by evolutionary psychologists is compatible with and perhaps made more plausible by a family of *dual processing* theories about the mental mechanisms underlying reasoning and decision making that have gained increasing prominence in recent years. [9] Though these theories differ from one another in many details, they all propose that reasoning and decision making are subserved by two quite different sorts of system. One system is fast, holistic, automatic, largely unconscious, and requires relatively little cognitive capacity. The other is relatively slow, rule based, more readily controlled, and requires significantly more cognitive capacity. Stanovich (1999) speculates that the former system is largely innate, emerged relatively early in human evolution and, as evolutionary psychologists suggest, has been shaped by natural selection to do a good job on problems like those that would have been important to our hominid forebears. The other system, by contrast, evolved more recently and, "although its overall function was no doubt fitness enhancing, it is the primary maximizer of *personal* utility" (Stanovich 1999 , 150). This newer system is more heavily influenced by culture and formal education and is often more adept at dealing with many of the problems posed by a modern, technologically advanced, and highly bureaucratized society.

Since the new system requires more cognitive capacity, is more influenced by culture and education, and does not get used automatically, Stanovich hypothesized that there might be significant individual differences in people's ability and inclination to use it. More specifically, he reasoned, people with higher cognitive capacity, as measured by instruments like the Scholastic Aptitude Test (SAT), and

with a cognitive style that emphasizes "epistemic self-regulation" should do better on tasks that the old system was not designed to handle. Stanovich agrees with the evolutionary psychologists that many of the tasks studied in the heuristics and biases

tradition fall into this category. In extensive studies of these tasks he has shown that, while the average performance on these tasks is indeed quite poor, there are *some* subjects who give the answer that the Standard Picture suggests is normatively correct on *many* of these problems, and these subjects typically have significantly higher SAT scores and score higher on tests designed to detect cognitive styles that include epistemic self-regulation.

If Stanovich and other dual process theorists are on the right track, then the unbridled optimism sometimes suggested by evolutionary psychologists is unwarranted, since most untutored people do indeed lack the capacity to deal with a wide range of problems that are important in a technological society. But the glum pessimism often associated with the heuristics and biases tradition is not warranted either. Since the fast, automatic, and evolutionarily older system requires little cognitive capacity, everyone has the capacity to deal rationally with many reasoning and decision making problems that were important in the environment in which we evolved. Moreover, since the new, slow, rule-based system can be significantly affected by education, there is reason to hope that better educational strategies will improve people's performance on those problems that the old system was not designed to deal with. This hope is encouraged by the findings of Nisbett (1993) and his colleagues showing that, on many sorts of reasoning problems, a little education goes a long way.

NOTES

Parts of this chapter are based on earlier work, especially Samuels, Stich, and Tremoulet 1999 ; Samuels, Stich, and Bishop 2002 ; Samuels, Stich, and Faucher 2002 ; and Samuels and Stich 2002 . We are grateful for the many valuable discussions, comments, and criticisms occasioned by this earlier work. Special thanks are due to Michael Bishop, Stephen Downes, Luc Faucher, Peter Carruthers, Richard Foley, Gerd Gigerenzer, Alvin Goldman, Daniel Kahneman, Ernie LePore, Ernest Sosa, and Patrice D. Tremoulet.

1. For useful surveys of this literature, see Nisbett and Ross 1980 ; Kahneman, Slovic, and Tversky 1982 ; Dawes 1988 ; Piatelli-Palmarini 1994 ; Sutherland 1994 ; and Baron 2001 .

2. Though the Standard Picture is widely accepted among philosophers, psychologists, and especially economists, it is of course not the only account of rationality that might be used to assess the quality of people's reasoning and decision making. (See Samuels, Stich, and Bishop 2002 and Samuels, Stich, and Faucher 2002 for discussions of some of the main alternatives.) Nor, for that matter, is the Standard Picture without its

end p.298

problems. First, as Goldman (1986 , 82) and Harman (1986 , chap. 2) have both pointed out, it is far from clear in what sense, if any, a normative principle of reasoning can be *derived from* the rules of logic, probability theory, or decision theory. Nor is it clear *which* rules of these formal theories our judgments and reasoning mechanisms must accord with in order to count as rational. Indeed, serious disagreements still persist over

which *versions* of logic, decision theory, and probability theory the correct principles of rationality ought to be derived from. (See, for example, Gigerenzer 1991a).
3. One of the most carefully developed of these challenges is Adler's (1984) argument aimed at showing that the results in Tversky and Kahneman's "feminist bank teller" experiment do not support the claim that subjects are committing a systematic reasoning error. Rather, Adler maintains, Gricean principles of *conversational implicature* explain why subjects tend to make the apparent error of ranking (h) (Linda is a bank teller and is active in the feminist movement) as more probable than (f) (Linda is a bank teller.).
Another important family of challenges argues that when interpreting data from an experiment on reasoning, advocates of the heuristics and biases program typically assume that there is a single best way of applying the norms of the Standard Picture to the experimental task. But this is not always the case. Gigerenzer 2000 , for example, argues that there are usually several different and equally legitimate ways in which the principles of statistics and probability can be applied to a given problem and that these can yield different answers—or in some cases no answer at all. If this is correct, then obviously we cannot conclude that subjects are being irrational simply because they do not give the answer that the experimenters prefer. For an extended discussion of these challenges, see Samuels, Stich, and Faucher 2002 .
4. While Kahneman and Tversky's rhetoric, and Gould's, suggests that untutored people have nothing but normatively defective heuristics or "shoddy software" with which to tackle problems dealing with probability, Piattelli-Palmarini goes on to make the even more flamboyant claim that this shoddy software is more likely to get the wrong answer than the right one: "We areblind not only to the extremes of probability but also to intermediate probabilities—from which one might well adduce that we are blind about probabilities. I would like to suggest a simple, general, probabilistic law: Any probabilistic intuition by anyone not specifically tutored in probability calculus has a greater than 50 percent chance of being wrong" (Piattelli-Palmarini 1994 , 131–32).
Despite passages like this, we think a case can be made that (3) is not really a central commitment of the heuristics and biases research tradition. For our defense of this claim, see Samuels, Stich, and Bishop 2002 .
5. The conjunction of the first two central theses of evolutionary psychology constitutes what might be called the *Massive Modularity Hypothesis*. For more on this hypothesis, see Samuels 1998 ; Samuels, Stich, and Tremoulet 1999 ; and Samuels 2000 .
6. Despite this evidence, the hypothesis remains very controversial. Many authors have proposed alternative hypotheses to explain the data, and in some cases they have supported these hypotheses with additional experimental evidence. Among the most prominent alternatives are the *pragmatic reasoning schemas* approach defended by Cheng, Holyoak, and their colleagues (Cheng and Holyoak 1985 and 1989 ; Cheng, Holyoak, Nisbett, and Oliver 1986) and Denise Cummins's proposal that we posses an innate, domain specific *deontic reasoning module* for drawing inferences about "permissions, obligations, prohibitions, promises, threats and warnings" (Cummins, 1996 , 166). Still other

hypotheses that purport to account for the content effects in selection tasks have been proposed by Oaksford and Chater (1994), Manktelow and Over (1995), Fodor (2000), and Sperber, Cara, and Girotto (1995).
7. In Samuels, Stich, and Bishop 2002 , we argue that, despite the rhetoric, (6) is not a central commitment of evolutionary psychologists who have studied reasoning.
8. For a much more systematic defense of the views offered in this section, see Samuels, Stich, and Bishop 2002 and Samuels, Stich, and Faucher 2002
9. Among those who advocate dual processing theories are Evans (1984 , 1989), Evans and Over (1996 and forthcoming), Sloman (1996), Klein (1998), and Stanovich (1999).
end p.300

chapter 16 GENDER AND RATIONALITY

Karen Jones

Reason and rationality have been the subject of feminist focus—much of it critical—right from the emergence of modern feminist thinking in the late seventeenth century (Astell [1701] 1970 , Masham 1705 , Wollstonecraft [1792] 1997). Nor is it surprising that feminist theorists should have turned their attention to reason and rationality, given that women's assumed deficiency in rationality has been used to justify their exclusion from full ethical standing and from full participation in education, politics, and the professions. Thus, addressing questions of gender and rationality came to seem central to the feminist liberatory project.

In contemporary feminist theory, the claim is often made that reason and rationality are "gendered" or that reason is "male" or "masculine." There might seem to be something inherently paradoxical in arguing—thus using reason and conforming to norms of rationality—that reason and rationality are male-biased and hence, at least from a feminist standpoint, suspect. As Haslanger writes: "I should also note that to my mind, there is something peculiar about engaging in discussion and reasoned debate over the value, or legitimacy, or reality, of reason and rationality. If there is something wrong with our commitments to reason, I doubt we'll find it this way (and I don't know what we could do about it if we did)" ([1993] 2002 , 211).

Richards (1980) assumes that to critique rationality is to embrace a rhetorically dangerous and potentially self-undermining position: just as feminists cannot object to women's subordination while rejecting justice as a male-biased ideal, they cannot rebut sexist arguments while rejecting reason. Those in weaker positions should be especially wary of suggesting, as some feminists have (Greer 1970 , 108–9), that rational argument simply functions to serve the interests of the stronger.

However, if, following Rawls's distinction between the concept of justice and conceptions of justice (1971, 5–6), we distinguish between the *concept* of rationality and various substantive *conceptions* of rationality, we can do much to dispel this air of paradox while at the same time providing a clearer way of stating the positions of those theorists, including Richards, who reject feminist critiques of reason.

The term "rational" functions as a term of approbation with paradigmatic application to strategies for belief formation and action choice, and to the beliefs and actions so chosen.

Derivatively, the term applies to persons insofar as they conduct their practical and epistemic lives in accordance with the right norms for action choice and belief formation and insofar as they instantiate the character traits and virtues—whatever they are—that make them wise agents and responsible inquirers. Rational agents respond appropriately to their reasons, whether practical or theoretical.

The concept of rationality is thus fundamentally *normative.* It is in virtue of their recognizing this normative role of rationality that theorists with very different understandings of what the right norms for action and belief are and what character traits and virtues are required in order to respond appropriately to one's reasons (i.e., theorists with very different *conceptions* of rationality) can nonetheless be said to be offering rival ways of explicating the same, fundamentally normative, notion. Feminist critiques of reason and rationality can be understood, nonparadoxically, as critiques of particular (often partial) conceptions of rationality—critiques typically, though not exclusively, undertaken on the grounds that that conception does not after all enable agents to respond appropriately to their reasons.

Feminist stances toward gender and rationality divide into three broad camps: the "classical feminist" stance, according to which what needs to be challenged are not available norms and ideals of rationality, but rather the supposition that women are unable to meet them; the "different voice" stance, which challenges available conceptions of rationality as either incomplete or accorded an inflated importance; and the "strong critical" stance, which finds fault with the norms and ideals themselves, or with some subset of them. A variety of very different research projects, separated as much by what they presuppose about gender as by what they argue about rationality, are being pursued within this third critical camp and it will be the chief focus of this essay, but any adequate survey of the terrain requires that the other approaches be presented at least in outline.

end p.302

I. The Classical Feminist Project

Theorists pursuing this research project differ in their conceptions of rationality, but they are united in accepting the following three claims.

1. Available philosophical or commonsensical conceptions of rationality and of reason represent genuinely human norms and ideals.
2. The list of norms and ideals contained within available conceptions of rationality and of reason are sufficiently complete. They are sufficiently comprehensive to enable agents who follow them to respond appropriately to their reasons, whether practical or theoretical. That is, they fulfill the internal normative function of a conception of rationality, namely that of prescribing for wise agents and responsible inquirers.
3. The external normative functions assigned to rationality and reason are unproblematic. This claim requires further explanation. In addition to the (internal) normative role of making prescriptions for agents and inquirers, our capacity for rationality has been thought to ground our status as persons worthy of equal moral standing, and to be that

1. Available philosophical or commonsensical conceptions of rationality and of reason represent genuinely human norms and ideals.
   which separates us from (nonhuman) animals; moreover, the most worthwhile and distinctively human lives have been thought to consist in developing and exercising our rational capacities. Thus, our capacity for rationality has been accorded a normative significance not accorded to other human capacities and has been privileged ahead of those other capacities.

The classical feminist project emerged first in the writings of Astell ([1701] 1970), Masham (1705), and Wollstonecraft ([1792] 1997) but continues to have advocates (Richards 1980 , Green 1995 , Pargetter 1986). Advocates of this approach respond to the charge that women are deficient in reasoning ability not by critiquing the yardstick against which women are held to fail to measure up, but rather by pointing out that some women do indeed surpass men in their abilities, and that shortcomings, where present, are the result of inadequate education. Thus Astell writes: “Can we Think and Argue Rationally about a Dress, an Intreague, an Estate? Why then not upon better Subjects? The way of Considering and Meditating justly is the same on all Occasions” ([1701] 1970 , 90, cited in Atherton [1993] 2002 , 32). And Wollstonecraft, somewhat tongue in cheek, compares women's reasoning ability with the equally deficient capacities of soldiers: “As proof that education gives this appearance of weakness to females, we may instance the example of military men, who are, like them, sent into the world before their minds have been stored with knowledge or fortified by principles.Soldiers, as well as women, practise the minor virtues with punctilious politeness” ([1792] 1997, 131). Astell, Masham, and Wollstonecraft each use their commitment to the central
end p.303

importance of human reasoning capacities (point 3 above) to ground powerful arguments for women's education. Only by being educated can women develop their capacities for rationality and hence become full moral agents.

II. The Different Voice Project

Proponents of this project, a project that has been largely pursued in connection with Gilligan's 1982 critique of Kohlberg's 1981 theory of moral development, deny claims 2 or 3 above (or both). That is, they claim that extant philosophical or commonsensical conceptions of reason and rationality are problematic not for what they include, but for what they leave out (McMillan 1982 , Ruddick 1980). Important capacities that enable us to respond well to our reasons are overlooked or devalued within traditional conceptions of rationality because those capacities have been associated with women and with women's nurturing activities within the private sphere, activities that have themselves been devalued. For example, some conceptions of rationality construe emotions as impediments to rational choice and to the reliable formation of true belief, but there is

evidence that our emotions enable us to respond to our reasons, both practical and theoretical (see Greenspan's contribution to this volume); thus the devaluation of those capacities traditionally viewed as feminine has resulted in a less than adequate conception of rationality (Jaggar 1996). It becomes a feminist goal to reinvest with value those capacities that have been associated with the feminine with a view to developing a more adequate conception of rationality that represents a genuinely *human* ideal.

III. The Strong Critical Project

Projects in this camp, while otherwise heterogeneous in philosophical commitments, are united in denying the first claim constitutive of the classical feminist approach. That is, they deny that available philosophical or commonsensical conceptions of reason and rationality represent genuine human ideals. The ideals themselves contain some sort of male bias. Most often this bias is explicated in terms of those ideals being "gendered," or "male," or "masculine," but, as we shall see, one might hold that the ideals or norms are biased without holding that
end p.304

they are gendered. For example, they might be biased in the sense that they function to maintain relations of dominance and subordination between the genders by deauthorizing women as rational inquirers.
Because the dispute among theorists in this camp is as much about gender as it is about rationality, it will be useful to classify theorists into rough family groups according to the conception of gender that underlies their claim that norms and ideals of rationality and reason are gendered. This will leave the position that the norms display some other sort of bias to be treated separately at the end of the section.

1. Gender as Symbolic

In *The Man of Reason* (Lloyd [1984] 1993), her pathbreaking study of ideals of reason in the history of philosophy, Lloyd argues that "rationality has been conceived as transcendence of the feminine; and the `feminine' itself has been partly constituted by its occurrence in this structure" (104). That is, our understanding of both rationality and femininity are shaped by this exclusion, so that we understand rationality as requiring the transcendence of traits associated with femininity and we understand femininity in terms of that which has to be transcended or left behind in order to become "the man of reason."
It is important to be clear about the claim being advanced here, lest it be subject to ready refutation (see Grimshaw 1986 , 61–69). Lloyd is not claiming that there is, in the history of Western philosophy, *a* conception of reason or rationality, and that there is, in Western culture, *a* conception of femininity and that these are in opposition with each other, such

that the traits and virtues ascribed to the ideal inquirer are incompatible with feminine traits and virtues. This would be to make a mistake about both gender and the history of philosophy: just as there have been many different conceptions of the traits and virtues of the responsible inquirer, so there have been many different conceptions of feminine traits and virtues. Rather, Lloyd is claiming that *whatever* the conception of reason, the feminine and all it stands for has been excluded from it, and that this exclusion has shaped our understanding of both what it is to be an ideal inquirer and what it is to be feminine.

Through readings attentive to metaphor as well as explicit argument, Lloyd identifies in the philosophical writings of thinkers as diverse as Augustine, Descartes, Hume, and Hegel, among others, a set of contrasts, including: reason/the senses, reason/the emotions, calm passions/regular passions, mind/body, culture/nature, knowledge/that which is known, universal/particular, theoretical/practical, and public/private. She argues that, either overtly in these texts, or covertly but revealed when a particular text is read in the context of the then-prevalent as

sumptions about gender, masculinity is aligned with the privileged and valued side of these contrasts and femininity with that which is inferior, or more subtly, complementary. Thus these contrasts are not value neutral: masculinity comes to signify all that is good, valued, and truly human, while femininity comes to be associated with that which is marginal, or not truly human, or apt to lead us astray. Moreover, these contrasts become available to symbolize the appropriate relationship between the sexes: just as reason should, in a well-ordered life, control the emotions, men should, in a well-ordered society, control women.

Suppose we accept that historical conceptions of rationality have indeed been deeply inflected by gender metaphors, what follows from this? Lloyd does not claim that, just in virtue of being associated with traits that are thought of as masculine, our norms of rationality are thereby at fault qua norms of rationality. That is, her critique of these norms is not *epistemic*: it is a further question whether, for example, a conception of rationality that opposes reason to the emotions enables us to respond appropriately to our practical and theoretical reasons. But this might lead one to wonder how deep the critique is and why either philosophers or feminists should be concerned about it: perhaps "it mistakes for real features of reason what are in fact mere superficial accretions of metaphor in its philosophical articulation" (Lloyd [1984] 1993 , 74). Why should feminists be concerned with the various associations that have come to attach to the concept of "the feminine"—what, we might wonder, has any of this got to with actual women and with the feminist liberatory project?

Even though there is no straightforward connection between the operation of "femininity" and "masculinity," or "male" and "female," as symbols in philosophical (or other) texts and the gender identities of actual men and women, the use of gender symbolism both is affected by and affects the social construction of gender identity (Lloyd [1984] 1993 , 74).

The use of gender metaphors in a text can also be crucial to a text's argumentative structure by making some inferences seem compelling and others not; thus, metaphor

should not be thought of as “mere.” [1] Sometimes metaphors bridge what would otherwise appear to be gaps in argument (as perhaps associations between femininity and emotions have delayed recognition of the ways in which emotions contribute to rationality). In addition, the use of gender metaphors can make such texts available to serve the ideological function of maintaining current gender relations. This can happen even though the text does not itself contain any false assertions, such as false assertions about the capacities of women or about the capacities needed to respond well to reasons. As a parallel, consider Darwin's theory of evolution. Darwin appropriated metaphors of competition and struggle originating in Malthus's population theory and used them to articulate the theory of natural selection. These metaphors compromised neither the truth nor the well-groundedness of Darwin's theory, but they contributed to making that theory available for (mis)appropriation by social Darwinians and sociobiologists. [2]
In these ways, gender metaphors can have flow-on effects both within a philosophical text and beyond it. Showing that a conception of rationality is gendered in the sense that it embeds gender metaphors in its articulation is not *sufficient* to show that it is inadequate as a conception of rationality—that is, that it makes a mistake about the norms, capacities and virtues needed to respond appropriately to reasons—though it rightly raises suspicions. Nevertheless, given the flow-on effects of such metaphors, they are of more than literary-critical interest and merit both philosophical and feminist scrutiny.

2. Gender as Psychosocial

A number of different projects explore the connections between psycho-development and theory construction with a view to defending the claim that particular philosophical theories or particular cognitive stances toward the world are masculine. Drawing on object-relations theory, Bordo offers a reading of Cartesian objectivity as a defensive response to “anxiety over the separation from the organic female universe” (1987 , 5). Flax 1983 finds that themes in Western philosophy—including conceptions of rationality—are to be explained as the working through of tensions in male psychosexual development, tensions theorized in the work of object-relations theorist Chodorow (see Chodorow 1977). Keller argues that the perception that science is a distinctively masculine enterprise might be explained by science's commitment to a particular conception of objectivity, a conception that postulates a sharp subject-object split in which the knower is seen as radically separate from that which is known, and nature is objectified (1985 , 79). This conception of objectivity, with its emphasis on separation and autonomy, answers to the concerns of male psychosexual development. Both male and female infants must develop a sense of their own autonomy and agency through separating from the mother—a process fraught with anxiety for infants of either sex—but, in social contexts in which an infant's primary caregiver is a woman, the male infant must also differentiate himself from the mother in order to establish his gender identity. This gives rise to psychological differences between the genders: boys are required to individuate themselves from the mother more radically than are girls (Chodorow 1977).

These projects have been subjected to various critiques: Bordo's reading of Descartes is contested (Atherton [1993] 2002). The approach of using object-relations theory to ground generalizations about the masculinity of philosophy has been charged with inappropriately generalizing about gender: the conception
end p.307

of gender that underlies this approach is alleged to be insufficiently historical and insufficiently attentive to the problem of trying to separate out the gender component from the complex identities of persons who are always also mem-bers of other socially marked categories (Grimshaw 1986 , chaps. 2 and 3; for a reply, see Bordo 1988). Even supposing that this account of gender does not generalize inappropriately, it focuses on gender difference rather than on gender domination (Bernick 1992 , 193). The critique of rationality to be examined next begins from an account of gender that focuses on relations of dominance and subordination.

3. Gender as Relational

Catherine MacKinnon writes:
The content of the feminist theory of knowledge begins with its criticism of the male point of view by criticizing the posture that has been taken as the stance of "the knower" in Western political thought.[That stance is] the neutral posture, which I will be calling objectivity—that is, the non-situated, distanced standpoint. I'm claiming that this is the male standpoint socially. I will argue that the relationship between objectivity as the stance from which the world is known and the world that is apprehended in this way is the relationship of objectification. Objectivity is the epistemological stance of which objectification is the social process, of which male-dominance is the acted-out social practice. That is, to look at the world objectively is to objectify it. The act of control, of which what I have described is the epistemological level, is itself eroticized under male supremacy. To say that women are sex objects is in this way redundant. Sexualized objectification is what defines women as sexual and as women under male supremacy. (MacKinnon 1987 , 50, cited in Haslanger [1993] 2002 , 227)
In highly compressed form, this passage contains MacKinnon's account of gender, her account of the targeted conception of rationality ("objectivity," "the non-situated, distanced standpoint"), and her account of the relationship between them. These elements of MacKinnon's theory and their interconnections are explored in Haslanger's "On Being Objective and Being Objectified" (Haslanger [1993] 2002), which offers a reconstruction of MacKinnon's argument for the claim that "objectivityis the male standpoint, socially."
First, the account of gender: according to MacKinnon, to be a woman is not to possess any particular set of psychological traits, nor is it to possess two X chromosomes or any other biologically defined trait or set of traits; rather, it is to have the relational and social property of being in a position of eroticized subordination: "Gender isa question of power, specifically of male supremacy

and female subordination" (MacKinnon 1987 , 40); "Male and female are created through the erotization of dominance and submission. The man/woman difference and the dominance/submission dynamic define each other" (MacKinnon 1989 , 113).

To be a woman is to be objectified—that is, to be seen and treated as an object for the satisfaction of (constructed) male sexual desire. To be a man is to occupy the social position of sexual objectifier. This social position of sexual objectifier is defined within power structures of dominance and subordination; thus being a sexual objectifier is not merely a matter of holding the objectifying attitude that women are *for* the satisfaction of male desire. Sexual objectifiers (men) have the power to make the sexually objectified (women) become who they see them as being (Haslanger [1993] 2002 , 225–26): "Men treat women as who they see women being"; "If woman is defined hierarchically so that the male idea of a woman defines womanhood, and if men have power, this idea becomes reality. It is therefore real. It is not just an illusion or a fantasy or a mistake. It becomes *embodied* because it is enforced (MacKinnon 1987 , 172, 119).

Gender norms—norms of masculinity and femininity—identify standards of excellence for persons in the relevant gender role (Haslanger [1993] 2002 , 213–16). This account of gender norms contrasts with gender-as-symbol accounts of femininity, inasmuch as no amount of symbolic association between a trait and women is sufficient to make that trait count as feminine. It also contrasts with accounts of gender as psychosocial: we need not suppose that women have the psychological traits identified as feminine by a theory of psychosexual development. Some women may have these traits, some may not, but failing to have those traits is insufficient to escape the gender identity "woman" so long as one lives within social structures of eroticized subordination, and so may be treated as a woman (Haslanger [1993] 2002 , 246 n. 23).

When gender roles are entrenched, the fundamentally prescriptive nature of gender norms is masked: people conform to those norms, and face sanction if they don't. Moreover, errant behavior is interpreted by those whose interpretations are taken to settle the matter through the lens of those norms, much as, for example, a slave's sabotage in tool-breaking is interpreted not as an expression of resistance, but as proof that slaves are "stupid" and suited only to menial labor. [3] According to MacKinnon, norms of objectivity have a role to play in this process: they lend support to the belief that women are by nature fitted to the socially subordinate position that they occupy. This belief, which projects onto women qualities that men desire that they have (e.g., docility and submissiveness), masks the fundamentally prescriptive nature of gender norms and thus lends stability to the oppressive relations constitutive of gender.

Haslanger calls the relevant norm of objectivity "Assumed Objectivity" ([1993] 2002 , 233) and analyzes it as composed of a cluster of four subnorms linking

observations to claims about a thing's nature and claims about a thing's nature to prescriptions about how to treat it. The norms are as follows:

- *epistemic neutrality*: take a “genuine” regularity in the behavior of something to be a consequence of its nature.
- *practical neutrality*: constrain your decision making (and so your action) to accommodate things' natures.
- *absolute aperspectivity*: count observed regularities as “genuine” regularities just in case: (1) the observations occur under normal circumstances (for example, by normal observers), (2) the observations are not conditioned by the observer's social positions, and (3) the observer has not influenced the behavior of the items under discussion. (Haslanger [1993] 2002 , 232–33)

But the norm of absolute aperspectivity would appear to be enough, if consistently followed, to alert an inquirer to the possibility that his own social position was affecting his observations. It would seem to be a way of weeding out projective beliefs, rather than lending support to them. To support the problematic projective belief, the norms require supplementation with:

- *assumed aperspectivity*: if a regularity is observed, then assume that (1) the circumstances are normal, (2) the observations are not conditioned by the observers' social position, and (3) the observer has not influenced the behavior of the items under observation. (Haslanger [1993] 2002 , 233)

Assumed aperspectivity in effect gives the inquirer epistemic carte blanche to ignore his social position and his influence on that which is observed. It, together with the three other norms of epistemic neutrality, practical neutrality, and absolute aperspectivity, jointly constitute the ideal of Assumed Objectivity.

In the context of gender relations (i.e., of relations of eroticized domination and subordination), following this set of norms will justify the projective belief that women are, by nature, as men have made them become, which in turn justifies treating women as women, with all the violence and indignity that MacKinnon claims that involves. We thus arrive at the conclusion that following these norms of objectivity, in current social circumstances, is functional to the maintenance of gender hierarchy, and thus that the norms are, at the least, male-biased. Insofar as the projective belief is false, since it mistakes for women's nature something that is the result of social processes of force, the norms also lead us into epistemic error and so are not appropriate norms for responsible inquirers (Langton 1993 , 381–84).

But we have not yet shown that the norms are gendered male; that is, we have not yet proven the claim that “objectivityis the male-standpoint, socially.” We must now investigate the connection between following those norms and functioning in the gender role of man—that is, in the role of sexual objectifier.

end p.310

Norms can be more or less tightly connected to the social roles for which they are norms: they can be norms that are *appropriate* to the role, or norms the following of which is sufficient to function in the role in question. Haslanger ([1993] 2002 , 220) offers the example of norms for tenants: being considerate of neighbors is a norm for tenants, but clearly you can satisfy this norm without being a tenant. Paying rent on time, in contrast, is a norm for tenants the satisfaction of which is sufficient to make you a tenant. Similarly, there are two senses in which a norm might be said to be gendered: a norm might be appropriate to the given gender role, or satisfying the norm might be sufficient to function in that gender role. Call a norm "weakly gendered" if satisfying that norm helps one achieve success in the relevant gender role; call a norm "strongly gendered" if satisfying that norm is enough to function in that social role (222). Haslanger concludes that the norms of objectivity are weakly (contextually) gendered inasmuch as satisfying them *in the context* of relations of gender hierarchy contributes to success as an objectifier through justifying both objectifying beliefs and action, but they are not strongly (contextually) gendered. Women may follow these norms without thereby functioning as men, but if they follow them they collaborate in their own objectification (235). [4]

It seems, then, that Haslanger has precisified the claim that rationality is gendered male in such a way as to enable a thorough evaluation of that claim. The evaluation of the claim proceeds on the basis of a specific conception of rationality, a conception given by the cluster of four norms that constitute Assumed Objectivity, and on the basis of a specific account of gender, an account that analyzes gender as a matter of social location in structures of eroticized dominance and subordination. The evaluation reveals a sense in which it is true that rationality is gendered male insofar as following the norms of Assumed Objectivity contributes to success as a sexual objectifier, but it does not support a strong claim identifying the standpoint of objectivity with the male standpoint. This precisification has, however, come at a price.

First, as Haslanger herself notes ([1993] 2002 , 240), it is hard to see the resulting critique as a critique of traditional philosophical (or even commonsensical) conceptions of rationality. Once exposed, no one would endorse the cluster of norms together called Assumed Objectivity. Moreover, as Langton notes: "What is bad about Assumed Objectivity is not that it is objective, therefore male, therefore oppressive, contrary to MacKinnon. What is bad about the norm is that in a sense it is *not objective enough*" (Langton 1993 , 382).

Indeed, the problem for the cluster of norms that constitute Assumed Objectivity emerges only when inquirers are given the epistemic license of *assumed aperspectivity,* and that norm, we might rightly think, is a norm that leads us away from, rather than toward, objectivity. We should attribute to MacKinnon a target conception of objectivity that lacks intuitive plausibility, lacks defenders, and identifies the problem as being one of *insufficient* objectivity when she identifies the problem as one of objectivity, only if a more plausible target for her critique cannot be found.

Second, there are aspects of MacKinnon's critique of objectivity that are inadequately captured by a reconstruction that brings critical focus to the norm of assumed aperspectivity. In particular, MacKinnon seems to find fault with (something like) Haslanger's norm of absolute aperspectivity and not merely with the unwarranted epistemic license conferred by assumed aperspectivity. MacKinnon seems to think that

aperspectivity disparages the insights of feminist consciousness raising that reveal how one's social location can *contributepositively* to one's epistemic standing (MacKinnon 1989 , 83–105).
In the next section, I explore an alternative reading of MacKinnon that focuses on the connections that MacKinnon draws between norms of objectivity and norms of credibility. This reading gives up on the project of arguing that norms of rationality are gendered male or are masculine; thus, it departs not only from Haslanger's reconstruction of MacKinnon, but also from symbolic and psychosocial accounts of how rationality might be gendered male. [5] The reading locates her critique of objectivity within a naturalist conception of how to defend norms and ideals that are norms and ideals for *us*—for the kinds of finite, embodied, socially located beings that we are.

4. Norms of Rationality and Norms of Credibility

Agents and inquirers who conform to norms and ideals of rationality are those whose claims to knowledge must be taken seriously. The concept of "rationality" thus functions to separate agents into those fit to participate in ongoing conversation about what to do and what to believe, and those whose opinion can be overlooked or ignored. Many feminist critiques of rationality begin from observations of the ways in which norms of rationality function to deauthorize women as knowers. Women are dismissed as unreliable informants regarding their own experiences (Code 1995 , 58–82), as too emotional and too enmeshed in particular concerns to view practical problems objectively, as unable to follow abstract argument, and so on. Feminist contributions to public debate are likewise frequently dismissed as "unobjective" and "biased."
Norms of rationality thus ground norms of credibility. It is easy to underestimate the epistemic importance of norms of credibility if one has a conception of ideal knowledge seeking as an *individual* enterprise. On this conception of ideal knowledge seeking, responsible inquirers seek, as much as possible, to limit their cognitive dependence on others. Real knowledge is knowledge gained firsthand through experience or through independent consideration of argument (Locke
end p.312

[1690] 1987 , I.iv.23). Reliance on testimony should be discharged, preferably through checking up on the content of that which is said, but, where this is impossible, through checking up on the credentials of the testifier. (And if this checking up requires further epistemic dependence, as it typically does when assessing expert credentials, then that dependence should in turn be discharged [Mackie 1969–70 , 254].)
This individualistic conception of ideal knowledge seeking is inappropriate for the kind of cognitively and temporally limited beings that we are: our knowledge is made possible through epistemic dependencies on others. We routinely rely on others for information that we cannot acquire ourselves, in both everyday and formal contexts of inquiry. In science, the theory dependence of method (Boyd 1983) embeds epistemic dependency in at least the following ways: in the construction of scientific instruments (e.g., an imaging

technique might presuppose results from physics, chemistry, and computer science, results that it is practically impossible—and not good scientific practice—for the medical experimenter using the technique to check), in controlling for artifacts of the experimental design, and in the selection of what rival hypotheses to test against (where these are identified using sociological indicators such as whether the hypothesis is advanced by an appropriately credentialed research team, published in a reputable refereed journal, and so on). The acquisition of knowledge can progress as fast as it does only because such acquisition is a social rather than an individual enterprise; moreover, some things are knowable by finite creatures like us only because of cognitive specialization and teamwork (Antony 1995 , 72–73; Hardwig 1985 ; Hardwig 1991). The norms and social mechanisms that grant credibility to some and withhold it from others thus become proper objects of epistemological, and not merely sociological, study (Schmitt 1994 , Goldman 1999). Feminists contribute to this epistemological project by drawing attention to the ways in which assignments of credibility both reflect and reinforce gender hierarchy.

MacKinnon writes:

Having power means, among other things, that when someone says, “This is how it is,” it is taken to be that way. When this happens in law, such a person is accorded what is called credibility. When that person is believed over another speaker, what is said becomes proof. Speaking socially, the beliefs of the powerful become proof in part because the world actually arranges itself to affirm what the powerful want to see.Beneath this, though, the world is not entirely the way the powerful say it is or want to believe it is. If it appears to be, it is because power constructs the appearance of reality by silencing the voices of the powerless, by excluding them from access to authoritative discourse. Powerlessness means that when you say “This is how it is,” it is *not* taken as being that way. (MacKinnon 1987 , 164) [6]

end p.313

Central to the feminist critique of rationality as functional to the maintenance of male dominance is the thought that extant norms of rationality—and in particular norms that determine who is and is not appropriately “objective”—underwrite discriminatory practices of credibility (Code 1991 , Code 1995 , Fricker 1998 , Jones 2002). The argument for this claim comes in two stages: first, it must be shown that norms of objectivity serve the ideological function of excluding women from “authoritative discourse”; second, the norms themselves must be found to be at fault, for if the fault lies merely in ideologically driven *mis*applications of the norms, then the solution would seem to lie in advocating a more rigorous adherence to, rather than rejection of, those norms.

MacKinnon claims that the stance of objectivity consists of two ideals: distance and aperspectivity: “To perceive reality accurately, one must be distant from what one is looking at and view it from no place and at no time in particular, hence from all places and times at once” (MacKinnon 1989 , 97). To the extent that a putative knowledge claim can be shown to be the product of the inquirer's social situation, that claim is undercut as knowledge: “If social knowledge can be interpreted in terms of the social determinants of the knower, it is caused. Therefore its truth value, in this definition of the tests for truth,

is undercut. If it has a time or place—or gender—it becomes doubtful because situated" (MacKinnon 1989 , 98).
MacKinnon claims that this requirement of aperspectivity is a "strategy of male hegemony" (MacKinnon 1982 , 537). The strategy has two faces, simultaneously legitimizing male claims to social knowledge and delegitimizing feminist claims to knowledge.
Both faces of this strategy are illustrated in the debates surrounding research into innate differences between the sexes. I will focus on arguments from sociobiology that purport to show that gender hierarchy is an evolutionary consequence of the differential reproductive investments of human males and females, because the dynamic is especially close to the surface in these debates. [7] Trivers (1972) argues that the parental investment of females in their offspring is necessarily higher than the parental investment of males. Because the female must gestate the fetus, birth the child, and suckle it, females are limited in the number of offspring that they can produce. Evolution will thus favor female reproductive strategies that ensure the quality of the offspring, including strategies aimed at gaining access to resources for child rearing. Male reproductive strategies face different constraints: a male can impregnate numerous females. Males who mate frequently and with many different females will leave more—potentially significantly more—copies of their genes in subsequent generations. This puts males in competition for reproductive access to females, who thus constitute a limiting resource for males. Barash claims these evolutionary facts explain male aggressiveness, competition, and promis
end p.314

cuity, as well as polygyny, and "female docility and sexual fussiness" (Barash 1979 , 89; see also Wilson 1978 .
This basic adaptionist argument is bolstered with contested research on fetal androgenization and selected descriptions of animal behavior, behavior that has itself been described in terms of the gendered characteristics that it then lends support to. Thus, for example, Barash describes hermaphrodite worms as "*active* and *aggressive* when seeking to discharge their sperm, *demure* and *discriminating* when their more valuable eggs are at stake" (1979, 50, my emphasis). A scattering of anthropological observations complete the standard sociobiological package.
Sociobiologists appeal to a fact/value distinction to deny a political agenda behind their research: "I worry that it will be misinterpreted and used as support for the continued oppression of women. My intent has been only to explore the evolutionary biology of male-female differences, not to espouse any particular social, political or ethical philosophy" (Barash 1979 , 89). Feminist objections to this research, in contrast, are seen as politically motivated and are explained away as the predictable consequence of feminist commitments. Feminists respond by pointing out the various ways in which socially available background beliefs shape sociobiological inferences: they lend support to the anthropological and animal studies that sociobiology calls on; they make the postulated psychological mechanisms look plausible even though many psychological mechanisms could have produced the hypothesized adaptive behavior under conditions of selection and rival psychological hypotheses are not ruled out. [8]

The socially available background beliefs that lend plausibility to sociobiology are themselves shaped by gender relations; thus, these beliefs are to be explained by features of the theorists' social situation. It is simply not true that sociobiology conforms to the ideal of "the ostensibly non-involved stance, the view from a distance and from no particular perspective, apparently transparent to its reality" (MacKinnon 1982 , 538). However, given that the beliefs that feminists contest are relatively entrenched, it will tend to be feminists and not sociobiologists who are called on to defend the presuppositions of their position (Antony [1993] 2002 , 132–34). Thus credibility is differentially apportioned between feminist and mainstream views on gender. [9]

We have completed the first stage of the argument for the conclusion that norms of rationality underwrite discriminatory credibility practices. That is, we have shown that in contexts of gender subordination (and race and class subordination—the point will generalize) norms of aperspectivity and distance will tend to be *used* ideologically. But the argument remains incomplete. We have not yet been given a reason to reject the norms of objectivity themselves, for the trouble seems to lie in their unfair application: feminists are held to standards that mainstream research does not meet. The problem lies in who is and who is not allowed to assume aperspectivity, rather than in aperspectivity itself. If this is the problem,

then perhaps the solution lies in holding mainstream research to the same standards as feminist research; indeed, the problem with research that reinforces gender relations seems to be that it is *not objective enough* (Antony [1993] 2002 , 208). The norm of aperspectivity could even be an important feminist tool, for all that has been said so far. Feminists can appeal to that norm to show that sexist science does not meet its own standards for objectivity.

However, MacKinnon seems to think that the problem runs deeper; in particular, the norms of distance and aperspectivity are incompatible with the method of feminist consciousness raising. (In Haslanger's terms: the problem is with absolute aperspectivity (or something like it) and not with assumed aperspectivity.) Feminists cannot appeal to the norm of aperspectivity to engage in "ideology-busting" for their own epistemic practice—which they assume to be good epistemic practice—will not stand scrutiny by its lights.

The norm of aperspectivity disparages knowledge claims that can be explained as the result of the inquirer's social location. Non-universal properties of an inquirer, such as social location, particular interests, and specific emotional responses are seen as distorting or "subjective" factors. Subjective factors should be controlled for and insulated from affecting the outcome of inquiry. Inquiry that does not adequately control for the influence of these factors is dismissed as "unobjective." The ideals of neutrality and distance likewise recommend insulating inquiry from potentially distorting influences such as political commitments.

Consciousness raising violated norms of aperspectivity, distance and neutrality. It was self-conscious in its commitment to political change. It was committed to "claim[ing] as valid the experience of women, the major content of which is the devalidation of women's experience" (MacKinnon 1989 , 116). While the formats of consciousness

raising groups—a grassroots phenomenon chiefly of the 1960s and 1970s—differed, they focused on recounting women's day-to-day experiences, especially of intimate relationships, and on their emotional responses to those experiences. They found this personal realm to be political insofar as it was the domain in which the power relations that constitute gender were constructed, maintained, and exercised. In women's often inchoate responses to their day-to-day experiences were found the resources with which to understand women's social position. These responses of, say, anger and discontent, responses that were by the lights of prevailing assumptions about gender roles *un*justified, were not defended as veridical by following a procedure that gave equal weight (or indeed any weight) to the hypothesis that they were symptoms of a merely personal discontent (Scheman 1980). Rather, they were *assumed* to be justified responses to at least some aspect of women's social position. This assumption was important, since without the body of feminist theory that emerged from consciousness raising, from feminist political action, and from subsequent academic work on gender, feminists would not have been able to rebut the then-prevailing interpretations of women's experience.

Given that this method starts out from a detailed examination of women's lived experience, an experience both available because of and constitutive of women's gender subordination, it finds social location to be an epistemic asset rather than a liability: “The process identifies the problem of women's subordination as a problem that can be accessed through women's consciousness, or lived knowing, of her situation. This implicitly posits that women's social being is in part constituted or at least can be known through women's lived-out view of themselves” (MacKinnon 1989 , 95–96).

Thus, *if* we believe that feminist consciousness raising gave rise to knowledge about women's social position, then we must reject aperspectivity, neutrality, and distance as norms for inquirers. [10] Those norms are rejected not merely on account of the fact that they can be misused and so function to maintain gender relations. Even fairly applied, they deauthorize feminist claims to knowledge insofar as they judge the method of consciousness raising illegitimate as a method of rational inquiry. Moreover, following these norms would prevent women from acquiring social knowledge; thus following these norms hinders rather than helps inquirers respond to their reasons. In this way, the conception of rationality is rejected as inadequate qua conception of rationality.

If consciousness raising enables women to learn the truth about gender relations not despite but *because* it starts from presuppositions, including most centrally the presupposition of the at least partial veridicality of women's emotional responses, then that must be because those presuppositions are *true.* Research agendas that begin from true presuppositions can further knowledge. This insight is not new to feminist theory: the theory dependence of method means that inquiry must always contain presuppositions. What matters is whether those presuppositions are true (Antony [1993] 2002).

Bringing the issues of credibility and of consciousness raising to the center of a reading of MacKinnon's critique of norms of objectivity makes that critique more substantive than it is on Haslanger's reading. She is not attacking norms that, once made explicit, most people would repudiate. The norms of aperspectivity, neutrality, and distance, though somewhat vague, are part of a commonsense conception of rationality that continues to hold currency. If MacKinnon is right in thinking that methods like

consciousness raising can give rise to knowledge, then those norms must be rejected, for they cannot explain the success of that method.
On this way of explaining the theoretical significance of feminist critiques of norms of rationality, they become part of the broader project of investigating what norms enable *us*—that is finite, social, epistemically dependent beings, with a specific cognitive architecture, functioning in particular environments—reliably to latch on to our reasons, both practical and theoretical. In focusing on the specifically political dimensions of credibility, especially as these involve gender, they contribute a distinctive perspective on the cluster of issues examined in social epistemology. In focusing on the epistemic importance of emotions, they join a
end p.317

growing group of theorists who recognize the noncognitive grounds of some of our cognitive achievements.

NOTES

1. See Lloyd (1993) 2002 and Lloyd 1998 for a discussion of the significance of metaphor in philosophy. The role of gender metaphor in philosophical texts is being explored by a number of theorists working in the continental tradition; important contributions include: Irigaray 1985a and 1985b , and LeDoeuff 1991 . See also Alcoff 1995 and Whitford 1988 . In a related project, feminist science scholars have carried out detailed case studies of the impact of gender metaphors on scientific inquiry; see especially Bleier 1986 and Keller and Longino 1996 .
2. Darwin is aware of the potentially misleading effect of these metaphors; he writes: “I should premise that I use the term Struggle for Existence in a large and metaphorical sense, including dependence of one being on another” ([1859] 1968 , 116). This comparison between feminist work on metaphor and rationality and critiques of metaphor in Darwin, is made by Sturr (1998).
3. Marlene Gorris's film *A Question of Silence* illustrates this power of defining away acts of resistance. See also Hoagland 1988 , chap. 1. It is a powerful theme in MacKinnon's work, about which I'll say more in the next section.
4. It is worth noting that showing that a norm is weakly gendered is not enough to condemn it, even given the moral bankruptcy of gender roles. Kindness might contribute to one's excellence in the role of master. The role of master should be abolished, but it does not follow that kindness should be (Haslanger [1993] 2002 , 219). However, following the norm of assumed objectivity is incompatible with the feminist commitment to social change insofar as it legitimizes existing gender relations.
5. Thus, there will be passages in MacKinnon's writing that square poorly with this account. But that is true of Haslanger's reconstruction as well. What matters is whether the alternative reading brings out interesting features of MacKinnon's view that are obscured on current readings. I will argue that it does.
6. It is perhaps telling that Langton (1993 , 369), following a suggestion of Haslanger, substitutes “proven” for “proof” in the sentence, “Speaking socially, the beliefs of the

powerful become proof in part because the world actually arranges itself to affirm what the powerful want to see." But that focuses on the truth making, rather than the silencing function, of power and thus elides the issue of credibility that is the central focus of this passage.
7. For an extended discussion of gender in sociobiology, with references to the larger literature, see Nelson 1990 . MacKinnon's own examples concern authority in social and legal contexts, including the ways in which such authority affects the definition of rape and makes the harms of pornography invisible.
8. For an overview of the issues here, see Gould 1981 , Kitcher 1985 .
9. Feminist critiques continue to be ignored in mainstream work on gender and
end p.318

sexuality. For example, LeVay's much cited and praised book on sexual orientation (1993) uses contested work on androgenization and contested sociobiological arguments without so much as mentioning the feminist critique of that work. Norms and sociological mechanisms for determining credibility and thus determining who must be taken to be part of the ongoing discussion regarding gender and sexuality are here clearly functioning to feminist disadvantage.
10. There is a large feminist literature on the epistemic value of social location; see especially Hartsock 1998 , Code 1995 , and Harding 1991 .
end p.319

chapter 17 RATIONALITY AND PERSONS

Carol Rovane

This chapter will explore eight related themes: (1) Persons are not merely rational, but possess full reflective rationality. (2) There is a single overarching normative requirement that rationality places on persons, which is to achieve overall rational unity within themselves. (3) Beings who possess full reflective rationality can enter into distinctively interpersonal relations, which involve efforts at rational influence from within the space of reasons. (4) A significant number of moral considerations speak in favor of defining the person as a reflective rational agent. (5) This definition of the person has led Locke and others to distinguish personal identity from animal identity. (6) Although it is a platitude that a person has special reason to be concerned for its own well-being, it is not obvious how best to account for that platitude. (7) Groups of human beings and, also, parts of human beings, might qualify as individual agents and, hence, as individual persons in their own rights. (8) There is a sense in which the normative requirements of rationality are not categorical but merely hypothetical.
end p.320

1. What Kind of Rationality Do Persons Possess?

The most minimal condition that something must meet in order to qualify as rational is that it be an appropriate object of what Daniel Dennett (1978) calls the "intentional stance." When we adopt the intentional stance toward something, we first attribute intentional attitudes to it and, then, we go on to predict the thing's behavior by working out what actions are rationally mandated by the attitudes we have attributed to it. Thus predictability via the intentional stance depends on conformity with the requirements of rationality. But it doesn't depend on understanding them. Indeed, if Dennett is right, it is appropriate to take the intentional stance toward things that couldn't possibly grasp the requirements of rationality, such as thermostats and pigeons. Other philosophers are not so liberal about what they are prepared to count as rational. Unlike Dennett, they insist that something cannot qualify as rational unless, in addition to conforming to the requirements of rationality, it also grasps those requirements and apprehends their normative force. In other words, a rational being must see that it ought to be rational. [1]
Let us call this kind of rationality full reflective rationality. It is the kind that persons possess. Unlike thermostats and pigeons, persons have the idea that they ought to be rational. And this makes for a further difference. Unlike thermostats and pigeons, persons can qualify as rational even when they don't conform to the requirements of rationality—or, equivalently, even when their behavior is not predictable via Dennett's intentional stance. Being reflective, persons can inspect their own thoughts and actions and evaluate the extent to which they do and don't conform to the requirements of rationality. When they don't conform, they can respond to such rational failure by engaging in self-criticism and efforts at self-improvement. These self-critical activities show that persons are committed to being rational, and it is by virtue of this commitment that they can qualify as rational even in the face of rational failure. [2]

2. What Are the Normative Requirements of Rationality?

Although persons can be said to grasp the requirements of rationality and apprehend their normative force, they do not typically have explicit and articulate knowledge of what those requirements are. This should not be surprising, since it is a matter of significant philosophical and psychological controversy just what those requirements are. The knowledge that persons have of them is probably like the implicit linguistic knowledge that Chomsky (1966) claims human beings are born with, even though they can't (prior to linguistic theorizing) articulate it. But, however implicit their knowledge of the requirements of rationality might be, persons must nevertheless have some conception of them. Otherwise, they could not engage in the kinds of critical activities they clearly do engage in whenever they evaluate their own and others' thoughts and actions as being more or less rational. What follows is a very rough and schematic account of the sort of thing that persons must implicitly have in mind, by way of normative requirements of rationality, whenever they do engage in critical evaluation of their own and one another's rational performance.
The most general normative requirement that rationality imposes on a person is the requirement to arrive at and act upon all-things-considered judgments. [3] These are

judgments about what it would be best for the person to do in the light of all of its beliefs, desires, and other attitudes. All-things-considered judgments are the outcome of a variety of rational activities that together comprise a person's deliberations, such as the following: resolving contradictions among one's beliefs, working out the implications of one's beliefs and other attitudes, ranking one's preferences in a transitive ordering. Each of these rational activities is directed at meeting a specific normative requirement of rationality. In the cases just mentioned, they are the requirements of consistency, closure, and transitivity of preferences, respectively. Deliberation involves many more rational activities, each of which is similarly directed at meeting some specific normative requirement of rationality. It is not important for the purposes of this chapter to identify all of these rational activities and the specific rational requirements they aim to meet. The important point is this: all of the rational activities that jointly comprise deliberation have a common purpose, which is to contribute to the overarching rational goal of arriving at and acting upon all-things-considered judgments. If this seems implausible, try to imagine what it would be like to arrive at all-things-considered judgments without satisfying the many other, more specific normative requirements of rationality. If a person refused to satisfy the requirement of consistency, for example, but persisted in believing a contradiction, then there might be no such thing as what it was best for the person to do in the light of all of its beliefs. One belief might direct the person to perform a certain action while its contrary directed the person not to perform it. Similar problems would arise if a person refused to satisfy the requirements of closure and transitivity of preferences. The person would be refusing to consider all things in the sense required for deliberation. Let us call the state that would be achieved if a person were to succeed in this endeavor of arriving at and acting upon all-things-considered judgments the state of *overall rational unity.* And let us note that there is one overarching normative requirement of rationality on persons that incorporates all

of the other, more specific requirements like consistency, closure, and so on, namely the requirement to achieve overall rational unity.

It is important to bear in mind that all of these normative requirements of rationality—both the overarching requirement to achieve overall rational unity and the other, more specific requirements like consistency—apply only to individual persons and not to groups of them. This can be seen from our critical reactions. If one person believes two contradictory propositions, then we are bound to regard this as a failure of rationality on that person's part. But if one person holds one belief while another person believes its contrary, then we are not bound to regard either person as guilty of a rational failure. For each of them might have reasoned correctly from its own point of view, by arriving at all-things-considered judgments that take all of its background attitudes into account.

Here is another way to put the point. Persons who disagree are not under the same kind of rational pressure to resolve their disagreements that individual persons are under to avoid inconsistencies among their attitudes. A person's commitment to being rational always requires it to resolve the latter, whereas a person's commitment to being rational does not always require it to resolve the former. In fact, it can happen that this commitment actually requires a person not to resolve the former. This happens when the person judges

that another person with whom it disagrees is definitely mistaken but, also, quite closed to changing their mind.
To sum up: the normative requirements of rationality define what it is for an individual person to be rational. But this definition does not describe the actual rational performances of persons. Rather, it articulates an ideal of which persons can and typically do fall short. Persons are nevertheless committed to satisfying this ideal, at least insofar as they are committed to being rational. And this is so despite the fact that persons don't typically have explicit or articulate knowledge of it.

3. What Kind of Sociality Follows upon Reflective Rationality?

Reflective rationality is essentially self-conscious. It may follow that it is a social achievement. Certainly, there have been many interesting attempts to show that this is so, including Hegel's master-slave dialectic, Mead's account of social self-consciousness, Wittgenstein's anti-private language argument, and Davidson's triangulation argument. [4] There is a common theme running through all of these
end p.323

arguments, which is that we cannot achieve full self-consciousness unless we are led, through concrete social interactions, to regard ourselves from another's point of view. If any of the arguments were to succeed, it would provide us with a powerful antiskeptical result. We would be in a position to conclude that the sort of reflective being who can raise worries about whether it is in touch with an external world necessarily *is* in touch with it, because it necessarily is in touch with other minds that populate that world.
Such antiskeptical arguments might appear to take us beyond the topic of this chapter, which is to explore the connection between persons and rationality. But the arguments are worth mentioning for the following reason. Many philosophers have taken for granted that we ought to define the person as a reflective rational being. If we were prepared to accept such a definition, and if, in addition, we accepted any of the above arguments as sound, we could conclude that persons are beings who are essentially social as well as rational. This is bound to have interesting ethical implications.
Unfortunately, it is not possible to assess these difficult arguments here. But it is possible to pursue a related, if less ambitious, course. Rather than ask, Why are reflective rational beings necessarily social? we can ask instead, What kind of sociality do reflective rational beings possess insofar as they are social? In answering this latter question, we shall see that the kind of sociality in question is the kind that characterizes certain *distinctively interpersonal* relations.
It should be noted first of all that there is a very abstract form of sociality that automatically follows upon reflective rationality. To be a reflective rational being is to have a conception of oneself as possessing a rational nature, and to conceive oneself in this way is to conceive oneself as belonging to a single kind that includes all and only those beings who share the same rational nature. This conception amounts to a form of sociality because it incorporates an understanding that all the members of the kind that it

picks out not only share a common nature but also know that they do—where this means that they know this *of* one another. Yet the way in which they know this of one another is so abstract that it barely qualifies as social knowledge. For it seems possible that they could have the knowledge without actually interacting with one another in the world. This, of course, is the very possibility that the arguments mentioned above are designed to rule out. They aim to show that the kind of self-conscious reflection that incorporates this abstract social knowledge that all reflective rational beings have of one another somehow presupposes that they also stand in concrete social relations in the world. Insofar as reflective rational beings ever do enter into concrete social relations with one another, they can bring this abstract social knowledge to bear in them. Their mutual knowledge of their common rational nature makes it possible for them to influence one another in rational ways, by offering reasons to one another. Whenever they do this, they leave it up to the other to consider the reasons
end p.324

on offer from its own point of view; they leave the other free to reject those reasons or to accept them and, possibly, to act upon them.

Like the description of rationality in the previous section, this description of rational influence is in certain respects idealized. Persons can be influenced in rational ways only insofar as they are implicitly committed to satisfying normative requirements of rationality—requirements that they inevitably fall short of satisfying and of which they have little explicit knowledge.

However, precisely because the account is idealized in these respects, it is compatible with the existence of a whole lot of rational failure in the course of (attempted) rational influence among persons. It should not be assumed, therefore, that rational modes of influence are rare, or that they are confined to rarefied contexts like the academy or courts of law in which the various interlocutors are explicitly committed to providing reasoned arguments for one another's consideration. The sloppy and ill-considered interchanges of everyday life are also efforts at rational influence in the sense at issue. What distinguishes them as rational modes of influence is not that they involve perfect conformity to the requirements of rationality. What so distinguishes them is that they invite others to embrace certain attitudes or undertake certain actions by indicating that there is some reason to do so. These invitations do not presuppose that reflective rational agents are perfectly rational. Those who offer reasons and those who receive them can alike be prone to mistakes, laziness, confusion, etc. All that is required in rational influence is that those to whom reasons are being offered have the power, if they so choose, to take those reasons into account in their deliberations. It should be easy to see, therefore, that rational influence is ubiquitous in interpersonal relations. We might even go so far as to say that rational influence is in play whenever persons talk to one another. (There is one exception that ought to be registered: bullshit. But it seems fair to say that this exception is entirely parasitic on the normal case of genuinely rational influence.) [5]

There is a crucial respect in which this account of rational influence is not at all idealized. It does not portray such influence as essentially respectful. The initial description provided above might make it appear that it is essentially respectful. For the description emphasized that when we offer reasons to others we leave it up to them to accept or reject

them. And this amounts to an acknowledgement that they are independent agents with their own points of view. However, this acknowledgement may fall well short of anything that most moral philosophers would want to call true respect for the freedom and autonomy of others. To see that this is so, consider the fact that lies and threats are species of rational influence. They are not like pushes and tugs that bring about their effects in purely physical ways that lie entirely outside the normative space of reasons. On the contrary. Liars offer their own—as it turns out unreliable—word as a reason for others to believe something—leaving it up to them to decide, from their own points of view, whether to believe the lies or not. Similarly, those who impose
end p.325

threats on others also leave it up to them to decide, from their own points of view, whether to comply with the threats or not. Of course, a credible threat must be backed by a promise to make the victim suffer in some way if the threat is not complied with. But such a promise is hardly like a push or a tug. It too is an attempt at influence from within the normative space of reasons.

Thus rational modes of influence are not always pure and good, nor are they always careful and cogent. What distinguishes them is that they are attempts to influence others from within the normative space of reasons, by exploiting and appealing to another's capacity for reflective rationality. [6]

4. Is Reflective Rationality a Defining Condition of Personhood?

The most powerful considerations that stand in favor of defining personhood in terms of reflective rationality are moral considerations.

One moral advantage that follows upon defining persons in this way is it allows us to make the following equation: something *is* a person just in case it is the sort of thing that can be treated *as* a person.

Here is why the equation follows. We noted above that rational modes of influence are ubiquitous in interpersonal relations. In fact, they are the distinguishing mark of all distinctively interpersonal relations. Whenever persons treat one another specifically *as persons*, they are engaging one another's points of view and interacting with them from within the space of reasons. We have just seen that this is so even when persons are disrespectful and abusive of one another. Even in such cases they are usually treating one another as persons as opposed to mere things, precisely because they are appealing to and exploiting their common rational nature. (Bullshit might appear to be an exception. Yet it seems that even in the course of bullshit persons are appealing to and exploiting their common rational nature, albeit without rational purpose.)

There is no doubt that to embrace this definition and equation is to take a somewhat exclusionary view of persons. We would deliberately exclude from the kind “person” anything that we can't treat in distinctively interpersonal ways—for example, fetuses, the severely insane, the irretrievably comatose, and the hopelessly senile. But if this seems unduly exclusive, consider that the definition is highly inclusive as well. It entails that if

we can treat something *as* a person, then it *is* a person. And herein lies its moral advantage. For, whenever we find we can engage something in distinctively interpersonal ways—for example, in conversa tion and argument—then we cannot deny that we are confronted with a person. Any attempt at such a denial would be a form of prejudice. Moreover, it would be a *hypocritical* form of prejudice, because one would be explicitly denying someone's personhood while at the same time implicitly acknowledging it through interpersonal engagement. So, for example, it was necessarily hypocritical of the slave owners of the American South to deny that their slaves were persons, given that they implicitly acknowledged their personhood whenever they talked to them (and, even more revealingly, when they passed laws against their education). In contrast, it would not be hypocritical to deny that fetuses, the irretrievably comatose, the severely insane, and the hopelessly senile are persons (note that it would be nonsensical to pass laws against educating them). In making such a denial one would not be depriving these human beings of ethical significance. They would remain objects of affection, concern, respect, legal rights, and so on. All one would be doing is registering that there is an important ethical kind to which they do not belong, namely the kind that can engage one another in rational ways and, thereby, treat one another specifically as persons.

So, one clear moral advantage of any view that makes reflective rationality a defining condition of personhood is that it helps to expose this kind of prejudice against persons, in which persons find it possible to treat certain others *as* persons and yet hypocritically deny that they *are* persons. It has obviously been of some real historical and political importance to expose and rectify this hypocritical form of prejudice against persons. [7]

The temptation to succumb to hypocritical prejudice against persons is greater when weighty moral and political consequences follow upon acknowledging their personhood—as happens in any context where it is granted that rights and obligations attach to persons. But, of course, it is precisely when such rights and obligations are at issue that it is morally imperative not to succumb to the temptation. And there is no doubt that the sense of personhood that is typically at issue in philosophical defenses of such rights and obligations is the sense given by the definition in terms of reflective rational agency. This is true of Kant's moral theory and of all theories belonging to the social contract tradition that includes Hobbes, Locke, Rousseau, and, more recently, John Rawls. [8] Not only do these theories place persons at the center of our moral focus, but the considerations they offer for putting them at the center also have directly to do with reflective rational agency. It is qua reflective rational agents that persons are capable of contemplating Kant's categorical imperative and, also, of framing and evaluating social contracts with others like them. All of these theories provide us with moral considerations in favor of defining persons in this way.

Locke ([1690b] 1987) offered some other moral considerations that favor the definition, which speak to us somewhat independently of the various theories of rights and obligations. The moral dimension of personhood that was most central for him was accountability. He wanted to make sense of the idea that a person

could stand before God on Judgment Day and be held accountable for all its past actions. It is easy to see that this idea presupposes that persons are reflective rational agents. If the

Last Judgment is to be meaningful, the person judged must be able to apprehend its evaluative force. And this requires that the person have the very sort of capacity for self-criticism that goes together with reflective rationality. In addition, Locke also held that the meaningfulness of the Last Judgment depends on another moral dimension of personhood besides accountability, namely prudential self-concern. He took for granted that the judgment must be backed by sanctions—the reward of eternal bliss or the punishment of eternal damnation. And these sanctions would be meaningless to the person being judged unless the person were prudentially concerned for its own well-being. Like accountability, this kind of prudential self-concern presupposes that persons are reflective rational agents. They must be reflective and rational enough to regard the threat of divine sanction as a reason for acting in one way rather than another.

To sum up: we have reason to define personhood in terms of reflective rational agency if we agree with Locke that persons are, by nature, accountable for their actions and concerned for their own well-being; likewise if we embrace any of the standard theories of rights and obligations that attach to persons; and likewise if we think that there is something important about defining and exposing hypocritical prejudice against persons, in which we find it possible to treat something as a person and yet deny that it is a person.

5. Locke's Distinction between Personal and Animal Identity

Locke ([1690b] 1987) famously argued that the condition of personal identity is not the same as the condition of animal identity, and in doing so he inaugurated a philosophical dispute that has continued to the present. He took the distinction between personal identity and animal identity to follow from the definition of a person as a reflective rational being. However, before considering Locke's argument, it is important to see that, in principle, his “animalist” opponents can also accept that definition.

Initially, it may seem that animalists cannot accept the definition. For, as we have already seen, the class of persons so defined does not coincide with the class of human beings (or, for that matter, any other suitably endowed animals). The reason why is that no human being possesses full reflective rationality at all stages

of its life, and some human beings do not possess it at any stage of their lives. But, according to animalists, it does not follow that a person is not a suitably endowed animal. On the contrary. As they see it, the capacity for reflective rational agency is part of the native biological endowment of human beings—though there is no reason why it couldn't belong to other animals as well. And they take it to follow that the identity of a person, qua thing with that capacity, is just the biological condition of identity for human beings and other suitably endowed animals. Thus, in their view, personhood is a status that suitably endowed animals achieve at certain times of their lives. But the thing that has the status—that is, the person—is always an animal whose life begins and ends with biological birth (or, perhaps, conception) and death. [9]

In contrast, Locke held that the life of a person does not begin and end with these biological events. The life of a person is the life of a reflective rational being and, according to him, such a life begins and ends with the person's self-conscious awareness of it. Much of the rationale for Locke's view derives from his preoccupation with a proper understanding of how a person can stand before God on Judgment Day and be held accountable for all of its past actions. We noted above that it is a necessary condition for such accountability that a person be a reflective rational agent who can understand the normative force of God's judgment. Locke also insisted that a person must have first personal or self-conscious knowledge of its own past actions as its own. That is, the person must be able to regard all of its past actions as "mine." Accordingly, when Locke considered the question of what God must call before Him when He calls all persons to be held to account, he reasoned that it was neither necessary nor sufficient that God should resurrect everyone's animal body. The only condition that need be satisfied is that each person stand before God with a consciousness of its past. Locke concluded that this must be the condition of personal identity, namely the condition in which there is a single, continuous consciousness.

In order to drive the point home—that the life of a person, qua reflective rational agent who can be held accountable for its own actions, is not the life of an animal—Locke offered a thought experiment about a prince and a cobbler. He asked us to imagine that the consciousnesses of the prince and the cobbler are switched each into the other's body, and to consider who would be who after the switch. After the switch there would be two persons, each of whom could coherently be held accountable for its past actions. But we could not keep track of their identities in biological terms. If we did, we should have to say that the person with the princely body and the cobbling consciousness continued to be the prince and so ought to be held accountable for the prince's past actions, even though the person had no awareness of any princely past at all. Locke regarded this as absurd. He took it to be obvious that the person who could and should be held accountable for that princely past would be the person who was con

end p.329

scious of that past as its own. And, in the case being imagined, this would be the person with the cobbling body into which the prince's consciousness had been switched. Contemporary neo-Lockeans have defended Locke's distinction by offering more thought experiments modeled on his about the prince and the cobbler. [10] One crucial innovation is that they have tended to exploit the other moral dimension of personhood that Locke took for granted—not personal accountability but, rather, prudential self-concern. They typically ask us to imagine that our own consciousness—where that includes all of our psychological attitudes and characteristics, such as our memories, beliefs, desires, commitments, and so on—is going to be switched into someone else's body and theirs into ours. Then they tell us that one of the resulting persons will be made to suffer while the other will be given all that it most wants and cherishes. And, finally, they ask us to choose, on the basis of prudential self-concern, which person will receive which fate. Anyone who chooses that the person who gets their original consciousness and psychological characteristics should be spared the harms and given the goods (and that the person with their original body should suffer the harms and be denied the goods)

thereby shows that they endorse Locke's distinction between personal and animal identity.

Some have protested that Lockean-style thought experiments do not demonstrate anything more than a conceptual distinction between personal identity and animal identity. They maintain that it might turn out to be the case that, as a matter of scientific fact, a person's consciousness and psychological characteristics simply cannot be transferred from one human body (or brain) to another. [11] On this way of thinking, whether Locke was right to distinguish personal identity from animal identity is ultimately an empirical question. However, in the next two sections, we shall see that the idea of a person is, au fond, a normative notion. In a way, this has already begun to emerge. We saw in the previous section that the most powerful considerations in favor of defining personhood in terms of reflective rational agency are moral considerations. Locke himself was admirably clear on the point. Not only did he invoke moral concepts like accountability and self-concern in his arguments about personal identity, he also declared that the term "person" is a forensic term. Yet Locke certainly allowed that his normative starting point must eventually give way to a properly metaphysical investigation into the nature of persons qua things that have the normative significance that they do have. In our scientific age, it is natural to expect that such a metaphysical investigation might best be pursued by invoking the empirical methods of science. But we will see in the next two sections that there are ways in which persons, qua reflective rational agents, are not ordinary objects of metaphysical or scientific investigation, and that this is due to the way in which normative considerations are bound to impinge on any such investigation.

end p.330

6. Some Puzzles about the Rational Basis of Self-Concern

Prima facie, the recognition that some future state will be mine suffices to give me a special sort of reason to be concerned about it. We have seen that this normative assumption can hold of persons only insofar as they are reflective rational agents. We have also seen that the definition of persons as reflective rational agents does not suffice by itself—that is, without further supporting argument—to establish Locke's distinction between personal identity and animal identity. For animalists can agree that persons are reflective rational agents and yet deny that their lives can come apart from the life of a given animal in the way that Locke and neo-Lockeans have suggested. And, finally, we have noted that their dispute with Locke seems to be a purely metaphysical dispute that might be settled by future advances in science. However, the topic of prudential self-concern lands us with a normative dispute as well, concerning what would constitute an adequate rational basis for a person's prudential self-concern. We shall see in this section that it is very difficult to insulate the metaphysical dispute about personal identity from this normative dispute.

Both neo-Lockeans and animalists can agree that any adequate account of personal identity must show why it is that a person has reason to be prudentially concerned for its own well-being. But they would have to show this in quite different ways.

On their side, neo-Lockeans can argue that a person, qua rational agent, cannot fail to be concerned for its well-being for the following reason: the well-being of any rational agent will consist, in large part, in the satisfaction of its desires; and, so, since a rational agent is necessarily committed to acting so as to satisfy its desires it is, *eo ipso,* necessarily committed to acting so as to ensure its own well-being. As it stands, this account is not yet complete. For it has not explained why persons have reason to be concerned for their future well-being as well as their present well-being. After all, action always takes place in the present. And, so, it seems that the only desires that a person, qua rational agent, is necessarily committed to acting upon are its present desires. Yet neo-Lockeans can reasonably claim to find grounds for future-directed self-concern in their account of personal identity over time in terms of psychological continuity. Insofar as persons know that their lives will be characterized by such psychological continuity, they can frame and pursue long-term projects, including long-term relationships. In the context of such long-term projects and relationships, a person's present desires and activities aim to contribute to something that is temporally extended. And this means that the person's present concerns will be structured in a way that makes essential reference to its future concerns as well. [12]

Obviously, animalists cannot claim to ground future-directed self-concern in quite the same way. If I conceive myself in animalist terms, then I must recognize that is possible for my life to go on without significant psychological continuity of the sort that would permit me to carry on with my long-term endeavors. From the neo-Lockean point of view, it would not be irrational for me to disregard a future to which I am not psychologically tied. And this is a crucial normative ground on which neo-Lockeans can argue that I ought not to regard such a future as mine. I should not regard a future to which I am not psychologically tied as mine precisely because, without such a psychological connection, there is no rational basis for the kind of self-concern that characterizes the life of a person. [13]
However, animalists can also lay claim to a normative advantage over the neo-Lockean view, for they can insist that the neo-Lockean account of the rational basis of self-concern faces a difficulty that they can avoid. The difficulty is that the neo-Lockean account fails to preserve the sui generis character of prudential self-concern. Here is why self-concern is not sui generis in the neo-Lockean account just sketched: it portrays a person's future-directed self-concern as grounded in its commitments to long-term projects and relationships. Such projects are often shared with others, and relationships always are. So although it is true that our commitments to them may give us reason to be concerned for our own future well-being, it is also true that such commitments will give us reason to be concerned for the well-being of others who also participate in them. That makes our future-directed self-concern analogous to our concern for those others and, hence, not sui generis. Animalists seem to have a somewhat better prospect of giving an account that preserves the sui generis character of self-concern. They can posit a natural instinct for self-love that is not grounded in any of a person's substantive commitments or in any of the psychological facts that neo-Lockeans take to constitute personal identity. In their view, it might simply be a biological fact that human beings evolved in such a way as to have a natural concern for their own well-being that is different in kind from their concern for all other things. [14]

This is hardly the end of the matter. There are two ways in which neo-Lockeans can respond to the animalist argument just presented. First, they might question the rational pedigree of a biologically based instinct for self-love. Suppose that sustained critical reflection led us to see that the instinct was in tension with other values to which we are firmly committed. In that case, we might conclude that the instinct does not have an adequate rational basis after all, and that this undermines the alleged advantage of the animalist view. This first line of response can be backed by a second, which claims that there is actually an advantage in *not* portraying self-concern as sui generis, for to grant that there is an analogy between self-concern and concern for others constitutes a significant advance in our moral thinking.

It is generally agreed that Sidgwick (1907) was the first to argue in this fashion, though many others have followed suit. Often, such arguments begin by pre

end p.332

senting the analogy between self-concern and concern for others in the negative light cast by the "present aim theory of rationality." As that theory's name suggests, all that rationality ever requires of us is that we deliberate from and act upon our present desires. One reason for embracing the theory is that it refuses to exaggerate the requirements of rationality. Another reason derives from reflection on the way in which deliberation and action always take place in the present. Since deliberation and action always do take place in the present, it seems that the only desires that it would be irrational *not* to take as the basis of one's deliberations and actions are one's present desires. It seems to follow from the present aim theory that we have no more reason to act in the present for our future well-being than we have to act for the well-being of others. It may have occurred to the reader that the neo-Lockean account of future-directed concern sketched above has already provided a solution to this problem for the present aim theory. This is a point to which we will soon return. But first, let's briefly consider some other responses that derive from more directly moral preoccupations.

Sidgwick was not content to view the analogy between self-concern and other-concern in a negative light. He observed that almost everyone acknowledges the moral demands of prudence. And he urged that anyone who is prepared to extend their concern for their own well-being as prudence demands, beyond the present and into the future, ought also to be prepared to extend their concern beyond their own case to the case of others.

Thomas Nagel (1970) takes a more Kantian perspective on the analogy, which leads him to seek a common ground of both prudence and altruism in the very structure of practical reason. According to him, if there is a rational ground of future-directed concern, it follows upon the existence of objective reasons that are not confined to our present desires because they are "timeless." Similarly, if there is a rational ground of altruistic concern, it follows upon the existence of objective reasons that are not tied to our own points of view because they are "impersonal." Nagel's Kantian strategy is controversial because it requires us to abandon the present aim theory of rationality.

Derek Parfit (1984) pursues a strategy that is more similar to Sidgwick's. He presents the case of a Russian nobleman whose present desires revolve around certain radical political commitments. The nobleman anticipates that he will become conservative in old age, and he sees that if he were to take measures in the present to secure his future well-being,

they would require him to act against his present radical commitments. From both a phenomenological and an evaluative standpoint, there is a sense in which this nobleman finds his own future desires just as alien, and just as opposed to his own present well-being, as the desires of another might be. Although the present aim theory of rationality would license the nobleman to disregard his future desires, Parfit insists that he can do so only if he refuses to recognize any moral grounds for altruism. For, insofar as he acknowledges any moral pressure at all to be altruistically concerned for others, he must extend that concern to his own case. He is no more licensed to disregard
end p.333

his own future desires than he is licensed to disregard the desires of others. Thus, while Sidgwick emphasizes that altruism cannot be any worse off than prudence, Parfit emphasizes in turn that prudence cannot be any be worse off than altruism.

Earlier, we raised a worry about whether any account of self-concern that affords a close analogy with concern for others can be right, given that it fails to portray self-concern as sui generis. The philosophers we've just discussed are quite happy to find that self-concern cannot be portrayed as sui generis, because they prefer a moral vision that demotes self-centered concerns in favor of the impersonal and impartial concerns of morality. The neo-Lockean account sketched above does not share that moral vision. It assumes nothing more than the present aim theory of rationality and goes on to argue that a person's present commitments can ground concern for its future well-being. This will be so insofar as a person is committed to long-term projects and relationships, for such commitments are bound to enlarge the scope of its self-concern beyond the present to include the future as well. In similar fashion, commitments to shared projects and relationships can enlarge the scope of one's concerns beyond one's own case to include others as well. But this kind of concern for others is quite unlike the sort of altruistic concern that Sidgwick, Nagel, and Parfit have in mind. It is neither impersonal nor impartial. It extends only to the particular persons with whom one is practically engaged through shared projects and relationships.

When concern for others is partialist, it is much more like self-concern. So perhaps the worry that the neo-Lockean account fails to portray self-concern as sui generis is not so grave as it might, at first, have seemed. That all depends on just how close the analogy is, within a projects-based account of self-concern, between self-concern and concern for others. In the next section, we will consider the possibility that it is not an analogy at all but, really, the same thing. As we do so, we will continue to find ways in which metaphysical investigations into the person, qua reflective rational agent, cannot effectively be insulated from normative investigations.

7. Rational Unity as a Condition of Personal Identity

It might seem paradoxical or, worse, nonsensical to say that the analogy between self-concern and concern for others is not a mere analogy because they are really the same

thing. But there is a way of thinking about joint agency according to which it is neither paradoxical nor nonsensical. On this way of thinking, when
end p.334

human beings exercise their agency together in joint endeavors, they can thereby constitute themselves as a single group agent. When this happens, what we would ordinarily think of as mutual concern on the part of the human constituents of the group agent might better be thought of as self-concern on the part of the group agent itself. The possibility of such group agents who are composed of many human beings goes together with another, which is the possibility of multiple agents within a single human being. By the lights of the definition of a person as a reflective rational agent, these amount to the possibilities of group and multiple persons. [15] This section will explore some reasons for granting that group and multiple persons truly are possible.

We saw in section 2 that there is a single overarching normative requirement of rationality, which is the requirement to achieve overall rational unity by arriving at and acting upon all-things-considered judgments that take all of one's beliefs, desires, and other attitudes into account. The point at issue in this section concerns the boundaries within which overall rational unity is to be achieved. Since the requirement defines what it is for an individual person to be fully or ideally rational, these boundaries must, of course, be the boundaries that mark one person off from another. The animalist view of personal identity implies that rational unity ought to be achieved within the biological boundaries that mark one human being off from another. The neo-Lockean view implies that such unity ought to be achieved within the phenomenological boundaries that mark one consciousness off from one another. (In real life there is not much difference between these two views, since there tends to be a one-to-one correspondence between individual human beings and individual personal consciousnesses.) In this section, we are exploring some reasons for thinking that neither view is correct, because of the way in which human beings can exercise their native rational capacities in order to achieve different levels of rational unity within different boundaries. If a group of human beings managed to achieve overall rational unity at the level of the whole group, then the group itself would have satisfied the very normative requirement that defines what it is for an individual person to be rational. And, precisely because this is so, it could effectively be treated as an individual person in its own right. The same might occur within parts of human beings. Indeed, something very much like this occurs in certain dramatic cases of dissociative identity disorder, where several alter personalities emerge, each of which can be separately engaged in its own right.

The claim that there could be such group and multiple persons is bound to meet with some skepticism. Unfortunately, it is not possible to give a full defense of it here. What follows is merely a rough indication of the kinds of considerations that support it. [16]

Let us consider the group case first. It is well known that when human beings engage in group activities, their joint efforts can take on the characteristics of individual rationality. Think, for example, of marital partners who deliberate to
end p.335

gether about how to manage their homes and families and other joint concerns. They may in the course of such joint deliberations do as a pair all of the things that individuals characteristically do in order to be rational: they may pool their information, resolve conflicts between them, rank their preferences together, and even arrive at all-things-considered judgments together about what they should together think and do—where the "all" in question comprises all of their pooled deliberative considerations. The same can also happen in a less thoroughgoing way when colleagues coauthor papers, or when teams of scientists design and run experiments together, or when corporations set up and follow corporate plans. We tend to assume that such joint endeavors leave human beings intact as individual persons in their own rights. Insofar as that is so, it should be possible to engage those human beings separately in conversation, argument, and other distinctively interpersonal relations. But sometimes this is not possible. Sometimes marital partners won't speak for themselves. Their commitment to deliberating together is so thoroughgoing and so effective that everything they say and do reflects their joint deliberations and never reflects their separate points of view. The same can happen to coauthors, team members, and bureaucrats. The kind of case at issue here is not one in which human participants simply wish to give voice to the larger viewpoint of the groups to which they belong. Rather, it is a case in which the human constituents of the group are not committed to having separate viewpoints of their own. That is, these human beings are not committed to achieving overall rational unity separately within their individual lives. Yet it is not because they lack rational capacities. It is because those rational capacities are directed in a different way, so as to help fulfill a larger commitment on the part of a whole group to achieve overall rational unity within it.

If this seems implausible, just think about two different attitudes you might bring to a department meeting. You might bring your own separate viewpoint to the table with the aim of convincing your colleagues to agree with you. This attitude takes for granted the status of each colleague as an individual person in its own right with its own separate point of view. The attitude also *perpetuates* that status, for the effect of adopting it will be that you maintain the separateness of your point of view by deliberating on your own, with the aim of achieving rational unity just within your own self. Even when you are moved by what your colleagues say, the reason why is not that you want to resolve disagreements with them, or do anything else that would help you to achieve rational unity as a group. You will be moved by your colleagues only insofar as what they say bears on your personal project of achieving such unity by yourself—by showing you that you have internal reasons, from your own point of view, to accept what they are saying.

But you might bring a quite different attitude to a department meeting, one that would not perpetuate your separateness from your colleagues. You might bring to the table all you have thought of with respect to the issues the department faces, with a view to pooling your thoughts with your colleague's thoughts, so that you can together discover the all-things-considered significance of the whole group's thinking. If your colleagues do the same, then it won't be true that each of you is committed to achieving overall rational unity on your own; you will be jointly committed to achieving such unity within the department. And, for this reason, it will be possible for others to engage the department itself in conversation, argument, and other distinctively interpersonal relations. The department could be asked, for example, why did you do such and such, and there will be a coherent answer that reflects the department's joint deliberations. The point is not that

departments of philosophy typically have the commitment that would render them sufficiently unified to be engaged in this way. The point is only that it is possible. It is possible for human-size philosophy professors to undertake a commitment to achieve rational unity together. And, if they were to live up to that commitment, then the lines that divide one person off from another would have shifted. They would no longer follow the biological divisions that mark off different human animals, or the phenomenological divisions that mark off different centers of consciousness. They would follow nothing else than the commitment to rational unity that is characteristic of the individual person. To say that these lines can be redrawn in these ways is to say that the facts of personal identity are, quite literally, matters of choice.

A similar argument can be produced for the possibility of multiple persons within a single human body. In fact, the argument is really a generalization of the argument for the possibility of group persons. What the argument shows, in effect, is that *all* cases of rational unity should be modeled on the unity of a group. Thus, the thought is that rational unity doesn't *just happen* as the inevitable product of some natural process, such as the natural biological development of a human being. Rational unity is something that is *deliberately achieved* for the sake of some *further end.* There are things that a philosophy department can do as a unified group person that no human-size person can do on its own. And that may constitute a *reason* why such human-size persons might initially decide to pool their efforts in a joint endeavor. If they implement their decision, they no longer maintain separate rational points of view. So, what perpetuates the group person once it has been brought into existence is not separate commitments on the part of its human constituents; it is up to the group itself to maintain its existence by continuing to strive for overall rational unity within it. When we view the unity of a human-size person along these lines, we must see it as deliberately achieved for the sake of some further end that couldn't be achieved without it. The appropriate contrast here is with an impulsive human being who doesn't strive for rational unity—who doesn't deliberate at all but simply follows current desires unreflectively and uncritically. Since the capacity to deliberate belongs to human nature, perhaps it is fair to say that such a human being is acting against its nature. But that doesn't harm the present point, which is that when human beings do exercise their rational capacities, they are *generating* rational unity

through their intentional efforts. And it is part of this same point that these capacities can be directed at the achievement of rational unity within different boundaries. An initially impulsive human being might come to strive for rational unity within each day, or week, or month, or year, or even a whole lifetime. The last goal was celebrated by Plato as part of the just life and by Aristotle as part of the virtuous life. In a less high-minded way, we now typically pursue the project of living a unified human life for the sake of other more specific projects such as lifelong personal relationships (friendships, marriages, families) and, also, careers. But what needs to be emphasized is that these are *projects* and they are *optional.* It is possible for human beings to strive for much *less* rational unity than these projects require and still be striving for rational unity. And, sometimes, the result may be relatively independent spheres of rational unity with a significant degree of segregation.

Such segregation is evident in degree in the lives of many human beings whom we find it possible to treat for the most part as roughly human-size persons. We may find, for example, that when we visit the corporation our friend "becomes" a bureaucrat who cannot recognize the demands of friendship at all. What this means is that our friend's life actually takes up a bit less than the whole human being we are faced with, the rest of which literally belongs to the life of the corporation. We often describe this phenomenon as a kind of "role playing." But it can also be described as a fragmentation of the human being into relatively independent spheres of rational activity, so as to generate separate rational points of view that can be separately engaged. Of course, group endeavors do not necessarily result in such fragmentation; in principle, they can completely absorb the human lives that they involve (this may actually happen in the armed forces and in certain very intense marriages). But when a group endeavor does *not* completely absorb the human lives that it involves, there is a consequent split in those lives. And multiple persons can be conceived along just these lines. The only difference is that the separate rational points of view of multiple persons need not be imposed by involvements in group projects but, rather, by involvements in other sorts of projects that it is not possible for a single human being to pursue in a wholehearted and unified way. When a human being's projects are numerous, and when they have nothing to do with one another, this may make it pointless to strive to achieve overall rational unity within that human life. And it may be a rational response to let go of the commitment to achieving such overall rational unity within that human life and to strive instead for as many pockets of rational unity as are required for the pursuit of those relatively independent projects. So, just as a group person may dissolve itself for the sake of human-size projects that would otherwise have to be forsaken for the sake of the group's overall unity, a human-size person may dissolve itself for the sake of even smaller projects that would otherwise have to be forsaken for the sake of the human being's overall
end p.338

unity. In such conditions, we will find the emergence of multiple persons within that human being, each of which can be treated as a person in its own right.

The central idea of this section has been that the rational capacities of human beings can be exercised so as to achieve overall rational unity within different boundaries. If this idea is correct, then we ought not to think of the facts of personal identity as a metaphysical given. We ought to think of them instead as products of effort and will. As such products, persons would be things that exist *for reasons* in a sense that goes beyond merely having causes. They would be things that exist for the sake of further ends that their existence makes it possible to pursue.

8. Why the Requirements of Rationality Might Not Be Categorical

The suggestions of the preceding section are surprising and, indeed, revisionist—not only with respect to the condition of personal identity but, also, with respect to the nature of the requirements that rationality places on persons. It is natural to think of those

normative requirements as being categorical in their demands. But the burden of the last section is that this is not so. Before explaining why this is not so, let us see why it at least seems so.

Certainly, the command “Be rational” can speak only to rational beings. Conversely, rational beings are the sorts of things that, by nature, are responsive to that command. That is precisely what distinguishes them as rational beings to begin with, namely that they are responsive to the command. On the face of it, it is hard to see what other or further reason rational beings could possibly require in order to see the command as binding, than that they have a rational nature by virtue of which they can apprehend its normative force. Because this is so hard to see, it is hard to see how the requirements of rationality could possibly fail to be categorical in their demands.

Even bracketing the arguments of the preceding section about group and multiple persons, there are some conditions in which it seems that the normative requirements of rationality might be set aside. However, we shall see that these are not best viewed as conditions *on* which the requirement to be rational is itself conditional. And, so, they don't serve to transform the apparently categorical imperative to be rational into a merely hypothetical imperative instead.

end p.339

Consider, for example, Derek Parfit's (1984) discussion of the “irrationality pill.” He points out that taking such a pill might be the best way to avoid succumbing to a threat, since taking the pill would make us unresponsive to the rational force of the threat. Here is the hypothetical imperative on which we would be acting in such a case: If you can achieve the all-things-considered best outcome only by being irrational, then be irrational. This is irrationality in the service of rationality. And it hardly serves to bring out any interesting sense in which the end of being rational fails to be unconditional for rational beings. It would be far more significant if it could be shown that the commitment to being rational is conditional on having other more specific and substantive ends.

It is certainly imaginable that a rational being might prefer to forgo engaging in rational activities. And this is imaginable even if the rational being regarded forgoing rational activity as a kind of suicide (because it would be the end of its life as a rational being). Now, if all this is imaginable, it is also imaginable that such a rational being might decide, upon reflection, not to commit this form of suicide because it had discovered that there are things worth doing—values worth pursuing and projects worth carrying out—that would require it to engage in rational activity after all. Such a discovery might be described as the discovery of a hypothetical imperative to be rational: if you find that there are substantive ends worth pursuing, then engage in whatever rational activities are necessary for the rational pursuit of them. A rational being to whom this hypothetical imperative speaks does seem to be one for whom the value of rationality itself is a substantive value that does not stand apart from other values and, so, is just as conditional in character as any other value might be. Yet there is still a sense in which the value of rationality can and must stand apart, even in such a case. The activity by which one reflects upon and comes to embrace the other substantive values in the light of which one can find reason to go on as a rational agent is already a rational process. And this means that a commitment to being rational is already implicitly presupposed in the very

evaluative process that is supposed to provide the conditional reasons for the commitment. It should be noted that there is no other value or end, besides the end of being rational, that necessarily figures this way in the life of a rational agent.
These considerations do not amount to an ironclad argument to the effect that the normative requirements of rationality constitute a categorical imperative for rational beings. But they do strongly suggest that this is so. Why, then, do the arguments of the preceding section suggest otherwise, and how do they do so?
The lesson is not that there is some deep mistake in the general direction of thought just outlined above. It is undeniable that the commitment to being rational is a necessary background condition for having other commitments to more substantive ends. It is also undeniable that a rational being is bound to be responsive to the normative requirements of rationality, on pain of losing its status
end p.340

as a rational being altogether. But there is something that these thoughts fail to take into account, which is that in order to be rational, a rational being must achieve overall rational unity. Thus, the command "Be rational" is really equivalent to the command "Achieve overall rational unity." And we have seen that the rational abilities of human beings can be exercised so as to achieve overall rational unity within quite different boundaries—within a whole single human being, or within a group of human beings, or within parts of human beings. And no matter what size persons already exist, it is within their power to redraw the boundaries between them by striving to achieve rational unity within different boundaries, either larger or smaller than the ones that currently individuate them. Once persons recognize that they face these alternatives, they need reasons to prefer one over another. Obviously, the command "Achieve rational unity" will not suffice to provide these reasons, because that command will be satisfied no matter which alternative is chosen. So persons must appeal to something apart from the bare requirements of rationality. They need to appeal to further, substantive ends, which would give them reason to achieve some given level of rational unity. As we put it at the close of the last section, persons exist for the sake of ends, which their existence, as the size agents they are, makes it possible to pursue. It follows that, when persons are commanded by rationality to achieve overall rational unity within themselves, we ought to allow that there is an implicit reference to and conditionalization on the ends for the sake of which they exist. The imperative, in other words, is an imperative to achieve overall rational unity within yourself, provided that you exist for the sake of ends worth pursuing and, furthermore, provided that you do not find other ends that are more worthy of pursuit, that would require you to strive to achieve rational unity within some larger or smaller boundary instead.

NOTES

1. Descartes expressed this view in *The Discourse on the Method* (in Descartes 1984a). See also Davidson's "Rational Animals" in Davidson 2001 .

2. Defenses of such a “commitment”-based view of rationality (and, indeed, of mental attitudes generally) can be found in Bilgrami 2002 , Brandom 1994 , and Levi 1997 . It should be noted that such an idealized understanding of the normative requirements of rationality is somewhat controversial. On many accounts, there should be very little slippage between rational ideals and rational performance. This would include all accounts of the mind that portray mental attitudes as dispositions and all decision theories that take for granted “revealed preferences.”
3. This conception of rationality is so widespread that it would be impossible to list all of its prominent subscribers. It can be found in Aristotle's ethics and it is widely assumed in contemporary decision theory.
4. See Hegel 1977 , Mead 1913 , Wittgenstein 1953 , and Davidson 2001 's “Rational Animals.”
5. For a philosophical analysis of bullshit, see Frankfurt 1988 , chap. 10.
6. For a fuller account of the many forms that rational influence among persons can take, see Rovane 1998 , chap. 3.1.
7. I argued for this importance in Rovane 1998 , chap. 3.2.
8. See Kant 1785 , Hobbes 1968 , Rousseau 1968 , and Rawls 1971 .
9. For defenses of animalism, see Wiggins 1975 , McDowell 1999 , and Olson 1997 .
10. This literature is very large. Two prominent contributions are Parfit 1984 and Shoemaker 1984 .
11. This is the view expressed in Wilkes 1988 .
12. For an insightful discussion of the role of long-term planning in the deliberative lives of persons see Bratman 2000 .
13. This line of reasoning is rarely made explicit, but it is implicit in every neo-Lockean thought experiment about personal identity that invokes self-concern.
14. Something like this idea is manifest in virtually all of the British empiricist writings on morals—except, of course, that they make no reference to evolution.
15. Korsgaard 1989 raises these possibilities.
16. See Rovane 1998 for a more complete defense, as well as an explicit formulation of the analysis of personal identity that follows, which I call there the “normative” analysis of personal identity.
end p.342

chapter 18 RATIONALITY, LANGUAGE, AND THE PRINCIPLE OF CHARITY

Kirk. Ludwig

Making sense of the utterances and behaviour of others, even their most aberrant behaviour, requires us to find a great deal of reason and truth in them.
—Davidson 1984 , chap. 10, 153

1. Introduction

This chapter deals with the relations between language, thought, and rationality, and especially the role and status of assumptions about rationality in interpreting another's speech and assigning contents to her psychological attitudes—her beliefs, desires, intentions, and so on. Central to the discussion below will be the
end p.343

status, in particular, of the Principle of Charity, first introduced by W. V. Quine as a maxim of translation, "assertions start[l]ingly false on the face of them are likely to turn on hidden differences of language" (Quine 1960a , 58–60). Donald Davidson has advocated a stronger form of the principle, which enjoins as necessary for interpretation of another's speech the assumption that she is largely rational and has largely true beliefs (Davidson 1984 , 27, 136–37, 152–53, 159, 168–67, 196–97, 200–201).
The discussion will be organized around the following three questions:

What is the relation between rationality and thought?
What is the relation between rationality and language?
What is the relation between thought and language?

These questions are not independent. To possess a language is to be able to speak to another and to understand another's speech. One must therefore be an agent, something capable of acting, as opposed to merely moving or being moved, to possess a language. Language therefore presupposes thought and action. If rationality is a condition on thought and agency, as is widely (though not universally) assumed, then it is likewise a condition on possessing a language. In this case, seeing another as a potential interlocutor carries a commitment to finding her to be fundamentally a rational being, and to regarding oneself as likewise fundamentally a rational being. On the other hand, there is a long tradition that sees language as essential for rationality, but not for thought generally. Aristotle famously defined man as "the rational animal" (1984 : *Topics*, book 5, 132a22–132a27; *Nicomachean Ethics* 1.7.1097b22–1098a20; *De Anima*, 1.5.645b14). We occupy an even more privileged position if, as Davidson has controversially argued (Davidson 1984 , chap. 11; 2001, chap. 7), language is a condition on thought, and rationality is essential to both: then thought, rationality, and language are possessed altogether or not at all.
Section 2 takes up the relation of rationality to thought. Section 3 discusses the relation between rationality and the power of speech. Section 4 takes up the relation of thought to language. Section 5 summarizes the discussion.

2. Rationality and Thought

It is widely accepted that rationality is essential for thought. This section explains what this view comes to, the reasons for it in outline, and some of the objections that have been advanced against it. We begin with a brief characterization of what
end p.344

is meant by “thought” in this discussion, and then of the domain and requirements of rationality.

The term “thought” will be used to cover any psychological attitude with a propositional content. The term “propositional attitude” (coined by Bertrand Russell ([1922] 1961 ; [1918] 1985) will also be used interchangeably with “thought” in this sense. Central examples are beliefs and desires. *A*'s belief that he is handsome has as its content that he is handsome; his desire to be admired has as its content that he is admired. Propositional attitudes are individuated by their psychological modes and contents. Thus, different attitudes can have the same content if they are entertained in different modes: one may have a belief that one will get well, for example, as well as a desire to get well. Other examples of propositional attitudes are intending, hoping, fearing, considering, wishing, and doubting. These are the psychological states especially relevant to a discussion of rationality because their contents, being propositional, can bear logical and semantic relations to one another; for example, one propositional content can require or support, or be incompatible or inconsistent with, the truth of another.

From antiquity, the domain of rationality has been divided into the theoretical, having to do with the formation of belief, and the practical, having to do with the expression of agency (the terminology is due to Aristotle; for contemporary discussions see Audi 2001 and Harman 1999 , chap. 1). Theoretical rationality aims at arriving at true belief and avoiding false belief, nonhaphazardly. Practical rationality, to which theoretical rationality is an important aid, aims at getting what one most wants, in accordance with one's beliefs about what one can get and how one can get it (and, perhaps, though controversially, with evaluating one's ultimate ends; see, e.g., Audi 1990a and Brandt 1998). The degree to which someone is rational depends on the degree to which his attitudes exhibit patterns at and across times appropriate for ideal pursuit of his theoretical and practical goals.

Theoretical rationality is concerned with having representational states that exhibit coherence at a time in the sense particularly of not displaying patterns that frustrate the goal of having true beliefs and avoiding false ones. Thus, for example, consistency in what one believes is an obvious goal of full or ideal rationality. Recognized inconsistency is worse than unrecognized inconsistency, though, and in cases in which it is difficult to discover the inconsistency we do not ordinarily count someone as irrational. (Frege's failure to recognize the inconsistency of his axioms for arithmetic does not convict him of irrationality.) Similarly for holding beliefs that, in the light of one's evidence, are not likely to be jointly true. Theoretical rationality concerns also how new beliefs are acquired in the light of new evidence, and with reasoning from, or acquiring new beliefs in the light of, beliefs which one already has. In general, the goal is to acquire true beliefs about, or relevant to, what one is interested in and to avoid false beliefs. Often what rationality requires is thrown into clearer relief by its break
end p.345

downs. Thus, wishful thinking, believing something because you want it to be true, and arbitrary belief formation, believing for no good reason, are irrational, while apportioning belief to the degree of evidence is rational. (See chap. 1, this volume.)

Practical rationality is also concerned with patterns of attitudes relevant to action, centrally belief, desire, and intention, both at and across times. An intransitive preference ranking is an example of an irrational pattern among conative states, since it can lead systematically to the frustration of one's practical interests (see chap. 10, this volume). If you prefer *A* to *B*, *B* to *C*, but *C* to *A*, then in principle you can be led to trade something of value (a penny, for example) in an endless cycle to get *B* for *C*, *A* for *B*, *C* for *A* and then again *B* for *C*, and so on, each pairwise trade seeming rational, though the entire set is not. Practical rationality concerns also the effective coordination of desires and beliefs in the pursuit of one's ends, which requires that one recognize what are the best means to ends most preferred and then implement them. Doing what one does not judge best all things considered—weakness of the will—is a familiar breakdown of diachronic practical rationality (see chap. 13, this volume, and Davidson 1980 , chap. 2). Similarly, though there is no general requirement on consistency in what one desires, there is a requirement on consistency in what one intends or plans to do, since inconsistent plans (the result of desires put through the sieve of practical reasoning) cannot be conjointly carried out (Bratman 1987 , chap. 8).

Having the power of thought and action obviously does not require perfect rationality, whatever that could come to. Most of us are subject to all too familiar failings in both reasoning and acting. This gives point to seeing rationality as a normative requirement, as a standard by which to judge our thought and behavior. The question whether rationality is required for thought is whether something can be a thinking being without being *largely* rational, or, more generally, what the limits are on how irrational one can be and still be seen as capable of thought. Thus, the thesis that thought requires rationality can be put as the thesis that propositional attitudes can appear only in largely rational patterns, synchronic and diachronic. This is to say that the normative requirements of rationality, which tell us how we ought to reason, deliberate, and act, are also descriptive requirements on what it is to be a thinking being: we think and act largely as we ought, or we do not do so at all.

The case for rationality being a requirement on thought rests on reflection on the conditions under which it is appropriate to attribute to something the basic attitudes of belief and desire, which are the primitive ingredients of agency. (We will assume that beliefs and desires come together or not at all—that is, that all thinkers are agents. While this might be challenged, there will not be space to discuss it adequately here.)

Beliefs come only in appropriate patterns. To see something as having one belief requires seeing it as having many related beliefs (Davidson 1984 , chap. 14, 200; Davidson 2001 , chap. 7, 97–102; Stich 1983 , 53–60). We would not attribute to someone the belief that a gun was in the desk drawer except insofar as we see her as believing that guns are artifacts, fire bullets, and have barrels, that desks are solid, that drawers open, have space in them, and so on. These general beliefs are conditions on possessing the concepts that are involved in the particular belief in question and express basic relations that hold between those concepts and other concepts and conditions relevant to their application.

One must also typically have many beliefs about particulars to think a gun is in the desk drawer, which are supported by the general beliefs required to have the concepts, for example, that the desk is not alive, that it takes up space, that it is larger than the drawer, and so on.

Action, which is the expression of agency, is seen in the light of both belief and desire. Agents do things. We wave to friends, we write letters, we prove theorems, build houses, cross the street. For present purposes, we can remain neutral on what actions are. Candidates are bodily movements, construed broadly to include certain mental events (Davidson 1980 , chap. 1), and causal processes (Dretske 1988). Actions are the products of intentions, which are formed in the light of our beliefs and desires. Typically the intention is formed on the basis of a desire for an end and a belief about how to achieve it. This shows the action in a favorable light, as done for a reason. Action explanations are often telescoped. We cite only the end or a connected means-end belief. We say, “He stepped on the brakes to stop the car,” or “He thought she'd be impressed by flowers.” The sense that an explanation has been given, however, depends on inferring that he thought stepping on the brakes would stop the car, or that he wanted to impress her. Davidson has argued influentially that every action is rationalizable by (made reasonable in the light of) a belief-desire pair that reflects means-end reasoning (Davidson 1980 , chap. 1). This has been disputed on the grounds that some actions are done for their own sake, so that no belief that it conduces to a further end is required to provide its reason (Locke 1974 , Mele 1988 , Mele 2003b). In any case, each action can be represented as the correlate of the conclusion of a bit of practical reasoning about what it is best to do, and how to do it or what constitutes doing it. To see something as an agent then minimally requires seeing it as exhibiting coherent patterns of belief and a certain kind of reasonableness in acting.

While these observations show that some minimal level of coherence in thought, desire, and action is necessary for something to be an agent, they don't by themselves guarantee that an agent cannot have many inconsistent beliefs, reason mostly ineptly, or act mostly on reasons that are not best all things considered. Further support for the view that agents' attitudes as a whole must be seen as appearing in a largely rational pattern lies rather in reflection more generally on the conditions under which we are willing to treat something as an agent. We should expect some unclarity about where to leave off calling something end p.347

an agent. Like most natural language terms, “agent” is semantically vague. Our practice does not determine a precise cutoff point along dimensions of variation relevant to its application. However, the degree to which a system can be seen as rational is clearly a relevant dimension of variation for the term “agent.” The less coherence we find in the set of attitudes we are thinking of as potentially those of an agent, the less clear we are that what we are considering is a possible agent at all. Moreover, reflection on cases—for example, step-by-step increases in overall incoherence—shows that to see another as an agent at all requires finding a large degree of coherence in his outlook, interests, intentions, and behavior. This coherence is expressed in seeing how the attitudes attributed both make for a reasonable picture of things, from the agent's point of view, and make sense of her behavior as an expression of agency. We do not treat anything as

an agent unless we find a large degree of reason in what it thinks and does. This can escape our notice because we tend to focus on follies that we can identify only in the light of the vast background of reasonable action and belief that makes sense of people having the attitudes we take them to have. Irrationality, seen in this light, is a "perturbation of reason," not its absence (Davidson 2001 , chap. 7, 99). We identify an irrational action or attitude as one that is a departure from an otherwise largely rational pattern, a pattern that makes sense of the attribution of attitudes that depart from it.

While this is the majority view, it has been challenged, nominally, at least, on both empirical and conceptual grounds. Investigation shows that these challenges, in addition to being subject to internal criticisms, involve a mischaracterization of the thesis.

Thus some psychologists have argued that experimental results show that human beings are not in fact rational animals (Johnson-Laird and Wason 1977 , Nisbett and Ross 1980 , Tversky and Kahneman 1983). "Pace Aristotle," one author says, "it can be argued that irrational behavior is the norm not the exception" (Sutherland 1994 , vii). Here is an example. Experiments show that most college students do poorly on some versions of the selection task. Consider four cards, with "E," "C," "5," and "4" on their faces. Each card has a letter on one side and a number on the other. The task is to turn over the minimum number of cards to check whether, if a vowel is on one side, an odd number is on the other. Many subjects turn over the "E" card and the "5" card, but not the "4" card. The conditional is falsified if a card with a vowel on one side has an even number on the other, so the "E" and "4" cards should be turned over. These and other experiments have been alleged to show that *most* people reason *irrationally* on many *simple* reasoning tasks.

However, it is clear that these observations about mistakes on reasoning tasks (as understood by the experimenter) do not undermine the view that it is constitutive of the propositional attitudes that they occur in largely rational patterns. Identifying what subjects believe and want to do in the experimental situations

end p.348

itself requires seeing their attitudes appearing in characteristically rational patterns. Their failures to reason as well as they could are themselves identifiable only because we already see them as largely reasonable creatures, responding largely reasonably to the tasks set before them. At most, then, these experiments could show there is some standard of rationality most people fail to meet. But, as Ernest Sosa (1999) has noted, to infer from this that most people are irrational or not rational is like inferring from the fact that human eyesight falls short of some extrahuman standard that most of us can't see well or can't see at all.

Apart from this, the interpretations of these sorts of experiments have been challenged on the grounds that they often involve overly simple assumptions about how the experimental subjects understand their task, and fail to distinguish between what subjects can do (competence) and what they do in the circumstances (performance). For example, we often use sentences to convey more than is conveyed by their literal meaning. We understand a card shark who says, "If it has this mark on the back, it's an ace," to be implying also that *all* aces have this mark on the back. Otherwise the remark would not be to the point. This would explain why in the selection task subjects turn over the "5"

card. Inattention and nonsalience can help explain why many don't turn over the “4” card. Conditionals are typically used in *modus ponens* reasoning rather than *modus tollens* reasoning. Inattentive subjects then are apt to check inference potential first. However, they understand the mistake when it is explained to them. Thus, one must also distinguish between what one is capable of on reasoning tasks in principle and one's actual performance, which may be affected by a wide range of conditions. (See Cohen 1981 and Davidson 1980 , chap. 14, 270–73, for further discussion.)

In addition to psychologists' efforts to show we are not in fact as rational as we might suppose, some philosophers have argued that there are only “minimal” conceptual constraints on the irrationality of agents. In particular, under this heading, it has been argued that there is no set of inferences or inference forms that all thinking beings must endorse, no “`fixed bridgehead' of true and rational beliefs” (Hollis 1982 , 73). The view that there is a fixed bridgehead of true and rational beliefs has been attacked in particular by Stephen Stich (1990 , chap. 2), who follows Cherniak 1986 in rejecting it on the grounds that we can imagine a people whose feasibility ordering for inferences is inverted with respect to ours: inferences we find easy, they find hard, and vice versa. Again, however, prima facie this is not a challenge to the thesis that propositional attitudes appear only in largely rational patterns. Nothing in the thought experiment suggests that the imagined people have massively inconsistent or incoherent beliefs or suffer from significant breakdowns of practical reasoning. In addition, the thought experiment that drives the argument is of doubtful coherence. The main difficulty is that the hypothetical others must possess the same concepts we do, for the complex inferences they are to find easy and we hard are

end p.349

ones couched in terms of our concepts. But we do not attribute concepts to people who are not able to recognize the simplest of inferences they conceptually underwrite involving them. So anyone who possesses those concepts must be able to recognize the validity of the same simple inferences we do (Biro and Ludwig 1994).

3. Language and Rationality

Given that agency and thought are necessary for language, if rationality is a condition on agency and thought, it is a condition on possessing a language. This section discusses how rationality is related to interpreting others as speakers. The next section takes up the question of whether language is necessary for thought.

Rationality, Interpretation, and the Principle of Charity

Speakers are agents. Hence, to possess a language is to be at least as rational as agents must be in general. A constraint then on interpreting another as a speaker, on assigning meanings to his sentences, and contents to his attitudes, is that the pattern of assignments makes him largely rational. Moreover, the attitudes attributed must make sense of him as an agent capable of performing speech acts and communicating with others on a potentially limitless range of topics. Because seeing another as rational is a matter of

finding his attitudes appropriately related by content and mode, this imposes a *holistic constraint* on interpretation, in the following sense. In finding patterns in another's behavior appropriate for interpretation as linguistic and other intentional behavior, one must be sensitive to the full pattern of assignments of meanings to sentences and attitudes in judging the appropriateness of the interpretation of any given bit of behavior.
Radical interpretation is interpretation of another without the usual contingent assumptions of commonality of language, culture, or psychology (Davidson 1984 , chap. 9). Reflection on radical interpretation is a tool in investigating the most general and abstract requirements on having a language. When we strip away all the common aids to interpretation and consider how we could come to interpret another simply on the basis of whatever we can know a priori about speakers, and on behavioral evidence open to public observation, we uncover what patterns in behavior we must perforce think are there if it is to be interpretable as linguistic behavior. This helps to show what content our linguistic and allied concepts
end p.350

have—those of meaning, truth, and the propositional attitudes—by showing how they are related to independent evidence for their application.
The requirement that we see others as largely rational in order to interpret them has most famously been discussed as a constraint on radical interpretation under the heading of "The Principle of Charity." We will concentrate here on Davidson's version of the principle. The principle has two distinguishable aspects, which are motivated differently. In his later work, Davidson has distinguished the two strands as the Principle of Coherence and the Principle of Correspondence (Davidson 1985a , 92; Davidson 2001 , chap. 14, 211).
The Principle of Coherence is concerned specifically with the a priori requirements on seeing something as an agent that can perform speech acts. It enjoins one to find another to be largely epistemically and practically rational, on pain simply of not being able to see the other as an agent at all. The grounding for this part of the Principle of Charity is a priori reflection on the nature of agency and the propositional attitudes of the sort sketched briefly in the previous section of this chapter. Davidson does not attempt to spell out in general the requirements on seeing others as largely rational, but argues that in practice it is a matter of seeing another as rational and reasonable by one's own lights, adjusting for differences in position and interests. This is consonant with seeing the principle as appealing to a priori constraints. For to say something is rational by our lights is just to say that we see it as required of a rational agent given our concept, that is, *the* concept, of rationality. Thinking about rationality from the point of view of interpreting another (or seeing another as an agent), in the light of the other's behavior, sheds some additional light on the concepts of agency and of the propositional attitudes, for it helps to highlight what gives them their practical content, namely, the way in which they enable us to make systematic sense of the movements of an object, by seeing it as goal directed in the light of largely true beliefs about the environment. To the extent to which we *cannot* make good sense of a system in these terms, we should not see it as being an agent with propositional attitudes: the supposition that it has propositional attitudes in this case is *idle*, and disconnected from the role the concepts of the attitudes play in our

making sense of things. (See Davidson's work in the philosophy of action, in particular the essays collected in Davidson 1980 , esp. chaps. 1–3, 5, 12–14; also Davidson 1982 and Davidson 2001 , chap. 7.)

The Principle of Correspondence has received three different formulations. To explain these and the motivation for the principle, it will be necessary to say more about Davidson's account of radical interpretation, for the Principle of Correspondence is introduced to solve a problem that faces the radical interpreter when his task is cast in a certain way. In particular, Davidson gives a central place in the interpreter's procedure to confirming for a speaker's language a formal recursive truth theory similar to the sort characterized by Tarski (1932) 1983 . The truth theory is to serve as the vehicle for a compositional meaning theory for the language (Ludwig 2002). The theory has as theorems sentences of the form (*T*) (or notational variants)

(*T*) *s* is true-in-*L* iff *p*

where "*s*" is replaced by a description of a sentence of the speaker's language in terms of its primitive significant parts, and "*p*" by a sentence in the interpreter's language. The radical interpreter aims to construct and confirm a truth theory that issues in theorems of this form by observing the speaker in his environment. Davidson has suggested that a truth theory confirmed from this standpoint can be used to interpret the speaker's utterances, that is, that we can use the sentence "*p*" used in the interpreter's language to give truth conditions for the speaker's sentence *s* in (*T*) to interpret *s*. This is in effect to hold that if a truth theory is confirmed from the radical interpreter's standpoint, "is true-in-*L* iff" in (*T*) can be replaced with "means that." (For natural languages, which contain context sensitive expressions, such as demonstratives and tensed verbs, the truth and meaning predicates must be relativized to features of context relevant to their interpretation.)

The radical interpreter must determine the appropriate (*T*)-sentences for the speaker's language. Davidson supposes the radical interpreter can, from behavioral evidence, figure out what sentences of his language a speaker thinks are true (holds true). The radical interpreter can then identify the conditions in the environment in which the speaker holds true a sentence. For some sentences, whether a speaker holds them true or not will not be sensitive to what goes on in his environment (e.g., "2+2=4"). In other cases, there will be systematic correlations between the speaker's holding true a sentence and what's happening in his environment (as in the case of "That's a rabbit"). These latter are the key to what the speaker means by his words and what he thinks.

A speaker's holding true a sentence is a "vector of two forces": what he believes and what he means by the sentence (Davidson 1984 , chap. 10, 196). Specifically, Davidson assumes (idealizing somewhat) that if a speaker believes that *p*, and believes that a sentence *s* of his expresses that *p* (at the time), then the speaker infers that *s* is true. For the speaker is committed to its being true that *p* and is presumed to know that if it is true that *p*, and *s* expresses that *p*, then *s* is true. Accepting this, if we knew which sentences a speaker held true and what they meant, we could infer what he believed. Likewise, if we knew what the beliefs were that formed the basis for his holding true various sentences, we could infer what the sentences meant. We start out, however, knowing neither what he believes nor what his sentences mean. The role of the Principle of Correspondence is to show how to break into this circle.

The first of the three formulations the Principle of Correspondence has received is that other speakers are largely in *agreement* with one, explicable error and ignorance aside. The second is that when a speaker holds true a sentence, by and large the sentence is true. The third is that a speaker's *beliefs*, particularly those that are responses to his environment, are largely true. Each has some textual support. However, the third is the correct formulation, for that is the only one that points to a way of fixing either what someone believes or the meanings of his sentences. For if we can assume that a speaker's beliefs about his environment are mostly true, then we know that when his beliefs are correlated with things in the environment, his beliefs are likely to be about things they are correlated with. If we further assume that there are enough constraints overall on interpretation to narrow down to a single salient condition, for each belief, what it can be reasonably correlated with for interpretation, we can identify what he believes on the basis of the correlated conditions that prompt his beliefs, for that is what they are about. If the third interpretation of the principle is correct, the correctness of the first two follows, but not vice versa.

The Principle of Correspondence is justified as a necessary condition on being able to construct a justified interpretation of another speaker on the basis of (i) a priori constraints imposed by our conceiving of him as an agent possessing a language, and (ii) information about his interaction with his environment. It plays a role similar to the Principle of the Uniformity of Nature, which holds that nature evolves in accordance with general laws, in Hume's account of what is necessary to justify our inductive inferences. We infer from past regularities to future regularities. For this to be reasonable, we must assume minimally that nature evolves according to general laws, which are reflected in past observed regularities, which can then be projected into the future.

Suppose that we knew that we could be justified in believing things about the future on the basis of the past. Then we would be able to infer that it is reasonable to accept the Principle of the Uniformity of Nature, since if it were not, we could not be justified in our inductive inferences. The Principle of Correspondence is justified in a manner similar to this, by appeal to the assumption that we can come to a justified interpretation of another speaker on the basis of the evidence and a priori constraints available in radical interpretation, and the claim that if we were not justified in believing the Principle of Correspondence, we could not come to a justified interpretation of another speaker on the basis of the available evidence, given the constraints.

The argument to justify the Principle of the Uniformity of Nature by appeal to our being justified in making inductive inferences is question begging. For we cannot know our inductive practices are reliable a priori, and any a posteriori justification of inductive practices must rely on the Principle of the Uniformity of Nature. Hume argued that the Principle of the Uniformity of Nature, since it

is itself a generalization that covers the past and future, and is not knowable a priori, could be justified only by appeal to induction, thus leaving both the principle and our inductive practices unjustified.

For the Principle of Correspondence to fare better than the Principle of the Uniformity of Nature, we must be able to know a priori that we can arrive at correct interpretations of any other speaker on the basis of radical interpretation. If we could know it only a posteriori, and the assumption of the Principle of Correspondence were necessary in any inference to what a speaker means by his words and what he believes, we could not support the Principle of Correspondence in this way without presupposing it.

Davidson adopts the strategy of justifying the Principle of Correspondence by appealing to an a priori justification of the possibility of correctly interpreting any other speaker from the standpoint of the radical interpreter. The argument for the possibility of correctly interpreting any other speaker from the standpoint of the radical interpreter rests on the observation that language is by its nature a medium for communication, so that interpretation must be something that can be accomplished on the basis of evidence that is available interpersonally: "That meanings are decipherable is not a matter of luck; public availability is a constitutive aspect of language" (Davidson 1990 , 314). The central idea here is that it is of the nature of a language that it is a device that enables its speakers to communicate with others, and that this can be done only on the basis of interpersonal evidence for what others mean.

Yet granting this does not seem to be enough to yield the result that Davidson needs. It would seem enough to satisfy the requirement that language be a medium for communication that it guarantee that *if* we had mostly true beliefs *and* were confronted with someone who speaks a language with largely overlapping expressive powers, *and* who believed similar things about the environment, we would be able to interpret him. Justifying the Principle of Correspondence requires the stronger claim that to have a language is to be interpretable correctly on the basis of public evidence in any circumstances by any possible speaker (see Lepore and Ludwig 2004, Ludwig 1992 , and Ludwig 1999 for further discussion).

Even if we despair of justifying the Principle of Correspondence, we need not despair of being able to interpret others correctly. For we need not suppose that our epistemic resources are restricted to those available to the radical interpreter. We can appeal to knowledge of features of our own psychological type and to the fact that in practice others whom we want to interpret are conspecifics embodied in the same way we are and in similar environments, to infer with some plausibility the sorts of things they are apt to be thinking, in order to constrain our interpretations. This would not, however, yield any argument to the conclusion that the empirical beliefs of linguistic beings were by their nature largely correct.

What is the relation between the two parts of the Principle of Charity? Traditionally, being rational has not been seen as requiring largely true empirical

beliefs. Thus, it has been supposed that one could be rational but mostly mistaken in one's empirical beliefs, someone, for example, who reasons perfectly, but who is systematically deceived about his environment by Descartes's Evil Demon, and, so, through no fault of his own. Rationality on this view is a precondition on having largely true empirical beliefs but does not require it. Rationality would require a large number of true general beliefs as a condition on possessing the concepts involved in any of an agent's beliefs, but

these would not be empirical. For example, to possess the concept of red, one would have to believe, indeed, to know, that red is a color, that red is a feature of the surface of an object, that no surface, viewed from one position, can be two different colors at the same time, that surfaces are extended, that extended objects occupy space, and so on. These propositions are not empirical propositions; they are necessary, and knowable a priori. If Davidson is right, one cannot be rational and a speaker without having mostly true empirical beliefs as well. To be a rational speaker, then, would be to be largely right about the world. Yet, while being largely right in one's empirical beliefs, if Davidson were right, would be necessary for being a rational speaker, it would not thereby be an aspect of being rational.

I have characterized the Principle of Coherence as being grounded in a priori reflection on the nature of agents. However, thinking about its role specifically in the light of interpretation of other speakers is useful in seeing why this is so. For if we think about our application of the concepts of the attitudes to other agents, it is clear that their utility lies in their enabling us to discern a pattern in the behavior of others that is usefully projectible. As Daniel Dennett (1987) puts it, adopting “the intentional stance” enables us to explain and predict the behavior of complicated systems in a way that would otherwise be practically impossible. Thus, contrast trying to predict where someone's body will be tomorrow at noon on the basis of its physical constitution and the laws of nature, on the one hand, with trying to predict where it will be on the basis of his intending to keep a lunch appointment with you tomorrow at the faculty club, on the other. It is the patterns among the attitudes and their relations to behavior imposed by the requirement that agents be largely rational that make for the practical utility of descriptions of objects in terms of propositional attitudes. That is, if a system has successfully been understood as an agent (and, hence, as rational) in the past, so that we have succeeded in interpreting its behavior as the product of rational agency, then, on the assumption that it continues to be an agent, we can usefully (if roughly) predict its future behavior from its present attitudes, any new attitudes it acquires, and what we suppose its reasoning powers to be. (Carl Hempel argued that it is the empirical assumption of rationality that enables us to predict agents' future behavior [Fetzer and Hempel 2001]; if being largely rational is essential to agency, the relevant empirical assumption is that the system is an agent. Of course, assumptions about the *degree* of an agent's rationality can still play a role in empirical explanation. In this regard, see also Davidson 1980 , chap. 14.)

end p.355

Language as Necessary for Rationality but Not for Thought

The discussion has assumed that rationality is constitutive of thought, and, hence, that it is constitutive of language. But, as noted in the introduction, it has been argued that while language is necessary for rationality, rationality is not necessary for thought. On the face of it, both views cannot be correct. However, if we consider what has been said in favor of the view that language is required for rationality, though rationality is not required for thought, we will see that proponents of the view have in mind stronger requirements on

rationality than those we have imposed above. Thus, these two views, though prima facie in conflict, are not actually so.

The view that language is necessary for rationality, but that rationality is not necessary for thought, has an ancient pedigree. Aristotle, and the Stoics, following him, held that while nonlinguistic animals had mental capacities, they did not have the capacity for reason, or, consequently, for belief (Sorabji 1996). This view was grounded on two claims. The first is that belief is assent or conviction as a result of persuasion (by others or oneself). The second is that reason is absent where there is no capacity for persuasion. It can be seen, however, that this is not so much a denial of anything we have said above as a use of a stronger conception of "belief" (*doxa*) or "reason" (*logos*) than we have been considering. For this is to think of *doxa* as what we would call belief arrived at on the basis of reflective reasoning, and *logos* as the capacity for reflective reasoning.

This brings out an important point about how "rationality" is now predominantly understood: not as a matter in the first instance of the capacity to engage in reflective reasoning but rather as a matter of thought and action being seen as reasonable in the light of evidence, belief, and desire, whether the agent reflects on these or not. Being rational, in the sense characterized in the previous section, and having the power of rational *reflection* are not the same.

More recently, Jonathan Bennett (1989 , 93) has argued that "possession of language is necessary for rationality" (though not sufficient), on the grounds that

(1)"Rationality requires the ability to manifest in behavior judgments about what is particular and past and what is general, that is, to manifest behavior that is appropriate or inappropriate to that which is not both particular and present";
(2)"Only linguistic behavior can be appropriate or inappropriate to that which is not both particular and present" (87).

Bennett explicitly denies, however, that "rationality" as it appears in this claim and argument "has anything to do with `rationality' in any contemporary sense of that term" (viii), and is rather intended "to stand for whatever human possession it is that creates a mentalistic difference of kind between us and other terrestrial animals" (vii, 4–5), that is, with giving a sense to "rational" that would truly make us the only rational animals. It is not surprising, given this stipulation, that it should turn out that only linguistic beings are rational. This thesis, then, like Aristotle's, and others in a similar vein, hinges on a conception of rationality that is not the one at issue above in the claim that rationality is necessary for any thought.

4. Language and Thought

The final question we take up is the relation between language and thought. If thought is prior to language in the sense that there can be thinking beings who are not linguistic beings, then the rationality of speakers is inherited (in part) from the rationality required

of agents. If language is required for thought, on the other hand, then there is a more intimate connection between having the power of speech and being rational. For then the power of rational thought and the power of speech would be interlocking capacities, each required for the other.

The majority view, unsurprisingly, is that thought does not require language. It is easy to see why this is so. First, it is natural and effective to adopt the intentional stance toward many animals. We explain the behavior of both domestic and feral animals by attributing to them beliefs about the world around them and desires similar to ours in basic respects. As Hume says, "Whenwe see other creatures, in millions of instances, perform like actions [to ours], and direct them to like ends, all our principles of reason and probability carry us with an invincible force to believe the existence of like causes" (Hume [1739] 1978 , 176). Second, we suppose ourselves to be continuous with the rest of the natural world, and it can seem incredible that human beings should represent the only evolved animals capable of even rudimentary thought. This seems especially incredible in the light of the natural view that language is a means of expressing thoughts that we antecedently possess. This idea is clearly expressed in book 3 of Locke's *Essay Concerning Human Understanding* ([1690b] 1987), which has had a profound effect on the development of thinking about the mind, even though it has long been recognized that Locke's theory is inadequate in its details. Given the ease with which we can see the nonlinguistic behavior of other animals as expressive of the same sorts of things that similar behavior in us expresses, the view that language merely adds the capacity to express the same basic types of psychological state can seem compelling.

There have been dissenters, however, from ancient times to the present, in different forms. Aristotle, as we've seen, held that animals were incapable of belief without language (though he seems to have had in mind more by *doxa* than we

end p.357

would require of belief). Plato held this too, though by virtue of holding implausibly that thought is silent speech; since Plato held that animals have thought, he attributed to them the power of language as well. Descartes famously denied nonlinguistic animals had minds (Descartes 1984a , 140–41; Descartes 1984b , 302–3; Malcolm [1973] 1991). Kant thought only the fully rational could have beliefs, and only those who possessed language could be fully rational. In contemporary philosophy, the independence of thought from language has been most famously challenged by Donald Davidson (Davidson 1984 , chap. 11; Davidson 2001 , chap. 7).

Davidson gives two arguments for the claim that only speakers have thoughts. The first is the argument from holism (cf. Stich 1979). The argument focuses on the kind of evidence available for the attribution of attitudes in the absence of verbal behavior.

The Argument from Holism

(A1)Beliefs and other attitudes are ascribable only in dense networks of attitudes.

(A2)Attributing a dense network of attitudes to an agent requires for support a rich pattern of behavior that gives substance to the attributions.

(A1)Beliefs and other attitudes are ascribable only in dense networks of attitudes.

(A3)The pattern of behavior required cannot be sustained in the absence of verbal behavior.

(A4)Therefore, only linguistic animals can have propositional attitudes.

Davidson recognizes that this argument is not conclusive. The main problem lies in sustaining premise (A3). The support comes from recognizing that there is something arbitrary about how we choose to describe the supposed attitudes of nonlinguistic animals. As Davidson says,

If we really can intelligibly ascribe single beliefs to a dog, we must be able to imagine how we would decide whether the dog has many other beliefs of the kind necessary for making sense of the first. It seems to me that no matter where we start, we very soon come to beliefs such that we have no idea at all how to tell whether a dog has them, and yet such that, without them, our confident first attribution looks shaky. (Davidson 2001 , chap. 7, 98)

A very complex pattern of behavior must be observed to justify the attribution of a single thought.I think there is such a pattern only if the agent has language. (100)

Thus, what is there to choose between saying that Rover thinks *a cat went up that tree* or saying that Rover thinks *a furred pointy-eared quadruped with whiskers*

end p.358

scurried up the oak he's barking up? Is there any further behavior that Rover could display that would decide which of these was more appropriate? It seems clear we should not take the particular sentence chosen for the belief attribution too seriously. How then can we take seriously the idea that Rover has any beliefs at all, for if he does have beliefs, must they not have determinate contents?

It is undeniable, however, that there is a pattern of activity that is captured equally well for practical purposes by either of our two attributions, but not by attributing to Rover the belief that the cat has vanished or turned into an acorn: Rover barks up the tree, peers into the branches, runs around its base, and finally settles down to wait (as we put it). (And there is the tale of Chrysippus's hunting dog, who, faced with three roads, checks the scent down two, before setting off down the last without checking.)

In light of this, perhaps we can characterize Rover's attitudes with classes of sentences that would do equally well for the purposes of accounting for his nonverbal behavior. Each particular sentence in the class attributes too much conceptual sophistication to Rover. To characterize his belief as captured equally well by any of them is, however, to characterize exactly the determinate content of his belief and the degree of refinement of his concepts, though we do not have words to express this in a single sentence. We can press into service Davidson's own favored analogy for disarming the threat of the indeterminacy of interpretation (Davidson 1984 , chap. 10, 154). Like the different scales we could use in measuring temperature, the different belief sentences we use to keep track of Rover's behavior (each of course requiring a supporting pattern of related

ascriptions) all equally well capture the phenomenon we are interested in, which is different from saying that there isn't any (Jeffrey 1985).
Davidson's primary argument, though, rests on claims about what is required to have beliefs, namely, the concept of belief, and what is, in turn, required to have the concept of belief, namely, language. The basic argument is as follows (Davidson 1984 , chap. 11, 170; Davidson 2001 , chap. 7, 102).

The Argument from the Concept of Belief

(B1)One can have propositional attitudes (thoughts) only if one has beliefs.

(B2)One can have beliefs only if one has the concept of belief.

(B3)One can have the concept of belief only if one has a language.

(B4)Therefore, one can have propositional attitudes (thoughts) only if one has a language.

The crucial premises are (B2) and (B3). Davidson offers an argument for (B2) in "Rational Animals" (Davidson 2001 , chap. 7, 104).
end p.359

The Argument from Surprise

(C1)"One cannot have a general stock of beliefs of the sort necessary for having any beliefs at all without being subject to surprises that involve beliefs about the correctness of one's own beliefs. Surprise about some things is a necessary and sufficient condition of thought in general."
(C2)"Surprise requires the concept of belief."

(C3)Therefore, one has beliefs only if one has the concept of belief.

(C2) seems true. Surprise requires recognition that something one thought was so is not, and so requires that one have the concept of belief. (C1) appears more susceptible to challenge, however. Davidson appeals bluntly to intuition, but this seems unconvincing. Child psychologists have argued that children (after initial language acquisition) pass through a developmental stage in which they cannot recognize that they had false beliefs (Gopnik 1990 ; Gopnik and Astington 1988 ; Perner, Leekam, and Wimmer 1987). If so, then despite possessing a language, and so thought, they are not capable of being

surprised. Whether this is right or not, it certainly seems possible. But if so, there can be no a priori requirement on thought or language that one be capable of being surprised. (C1) would furthermore lead to the surprising conclusion that there cannot be an omniscient being, since such a being is never surprised. However, we can secure (C3) by another route. For, granting that belief requires agency, having beliefs requires having beliefs about actions serving as means to ends (a belief that doing *this* is likely to bring about *that*), and so the concept of action. But the concept of an action involves in turn the concepts of belief, desire, and intention. Thus, one can have beliefs at all only if one has the concept of belief. Let us therefore grant (B2).

That leaves (B3). Davidson's main argument for (B3) can be reconstructed as follows (see Davidson 1984 , chap. 11; Davidson 2001 , chap. 3, chap. 7, 103–5, chap. 13, 202 for (D3), chap. 14):

The Argument from the Concept of Error

(D1) To have the concept of a belief, one must have the concept of error, or, what is the same thing, of objective truth, the contrast between how things are represented and how they are.

(D2) The claim that a creature possesses the concept of error, or objective truth, stands in need of grounding: this must take the form of giving an account of how there could be scope in the creature's experience for correct application of the concept.

end p.360

(D3) We can understand how there could be scope for the application of the concept of error in a creature's experience if, and only if, we conceive of it as a creature that is, has been, or is potentially (fixing the creature's capabilities) in communication with other creatures, and so able to use the concept of error as a tool in interpretation to achieve a better rational fit between a speaker's behavior and the beliefs and meanings we attribute to him; that is, the concept of error would have some work to do for interpreters of others' speech (there would be scope for its application in their practice), but not otherwise.

(D4) Therefore, from (D2)–(D3), to have the concept of error or objective truth one must be, or have been, or potentially be in communication with others.

(D5) Therefore, from (D1) and (D4), to have the concept of belief, one must be, have been in, or potentially be in communication with others.

Since communication with others requires one to have a language, (D5) suffices to establish (B3). Davidson nowhere states (D2) explicitly, but it appears to be in the background in the argument, which proceeds by asking after conditions that make sense of there being scope for the application of error in a creature's behavior or experience. Davidson states (D3) in a stronger form than given here, as requiring actual communication, but this seems too strong and is not required for the conclusion.

(D2) may be challenged on the grounds that all that is required for having a concept is that a creature has reason to think that there is scope for correct application of it in his experience. This is a familiar difficulty in transcendental arguments that rely on conditions for the possession of concepts to get to their correct application. However, even granting (D2), (D3) seems questionable. For while communication provides scope for the application of the concept of error, it is not clear why many other activities do not provide just as much scope for its application, such as, for example, correcting one's own past beliefs in the light of new evidence, or explaining behavior of a nonlinguistic animal that otherwise seems irrational in the light of a false belief, for example, explaining Rover's barking up the wrong tree by attributing to him the false belief that that is where the cat is.

Davidson's argument for the claim that language is essential to thought is therefore inconclusive. Consequently, we cannot support premise (2) in the following argument for the conclusion that language is necessary for rationality by way of Davidson's argument: (1) a being can think iff it is largely rational; (2) a being can think iff it possesses the power of speech; therefore, (3) a being is largely rational iff it possesses the power of speech. (See Heil 1992 , chap. 5, and Lepore and Ludwig 2004, chap. 22, for further discussion.)

5. Summary

Rationality is essential for thought and agency. It is constitutive of the propositional attitudes that they are attitudes of an agent, and nothing is an agent except insofar as its behavior can be interpretable as expressing largely rational behavior and thought. Since agency is a condition on language, rationality is essential for language as well. To say that thought requires rationality is not to say thinkers, and speakers, are perfectly rational, incapable of mistakes, or follies, neuroses, psychoses, and so on. Rather, what mistakes agents make, what irrational behaviors they engage in, are identifiable only against a background of largely rational thought and behavior, which makes sense of the attitudes involved in the irrational behavior or reasoning. Language may be argued to be essential for rationality only by employing a special standard for rationality, requiring advanced reasoning abilities, or by arguing that one can have propositional attitudes only if one is a speaker. Davidson's controversial arguments for the latter position are not conclusive. The ease with which we can explain animal behavior by attributions of propositional attitudes on the basis of the same sort of nonverbal behavior we often use with human beings suggests that thought, agency, and, hence, rationality, are not confined to linguistic beings. Language is not a precondition of rationality, but rather amplifies our powers of reasoning.

end p.362

chapter 19 RATIONALITY AND SCIENCE

Paul. Thagard

Are scientists rational? What would constitute scientific rationality? In the philosophy of science, these questions are usually discussed in the context of theory choice: What are the appropriate standards for evaluating scientific theories, and do scientists follow them? But there are many kinds of scientific reasoning besides theory choice, such as analyzing experimental data. Moreover, reasoning in science is sometimes practical, for example when scientists decide what research programs to pursue and what experiments to perform. Scientific rationality involves groups as well as individuals, for we can ask whether scientific communities are rational in their collective pursuit of the aims of science.

This chapter provides a review and assessment of central aspects of rationality in science. It deals first with the traditional question, What is the nature of the reasoning by which individual scientists accept and reject conflicting hypotheses? I will also discuss the nature of practical reason in science and then turn to the question of the nature of group rationality in science. The remainder of the chapter considers whether scientists are in fact rational, that is, whether they conform to normative standards of individual and group rationality. I consider various psychological and sociological factors that have been taken to undermine the rationality of science.

end p.363

What Is Science For?

First, however, it is necessary to deal with a prior issue: What are the goals of science? In general, rationality requires reasoning strategies that are effective for accomplishing goals, so discussion of the rationality of science must consider what science is supposed to accomplish. To begin, we can distinguish between the epistemic and the practical goals of science. Possible epistemic goals include truth, explanation, and empirical adequacy. Possible practical goals include increasing human welfare through technological advances. My view is that science has all of these goals, but let us consider some more extreme views.

Some philosophers have advocated the view that the primary epistemic aim of science is the achievement of truth and the avoidance of error (Goldman 1999). On this view, science is rational to the extent that the beliefs that it accumulates are true, and scientific reasoning is rational to the extent that it tends to produce true beliefs. The philosophical position of *scientific realism* maintains that science aims for true theories and to some extent accomplishes this aim, producing some theories that are at least approximately true. In contrast, the position of *antirealism* is that truth is not a concern of science. One of the most prominent antirealists is Bas van Fraassen (1980), who argues that science aims only for empirical adequacy: scientific theories should make predictions about observable phenomena but should not be construed as true or false. The antirealist view, however, is at odds with the practice and success of science (see Psillos 1999 for a systematic defense). Most scientists talk and act as if they are trying to figure out how the world actually works, not just attempting to make accurate predictions. Moreover, the impressive technological successes of science are utterly mysterious unless the scientific

theories that made them possible are at least approximately true. For example, my computer would not be processing this chapter unless there really are electrons moving through its silicon chips.

But truth is not the only goal of science. The most impressive accomplishments of science are not individual facts or even general laws, but broad theories that explain a great variety of phenomena. For example, in physics the theory of relativity and quantum theory each provide understanding of many phenomena, and in biology the theory of evolution and genetic theory have very broad application. Thus a major part of what scientists strive to do is to generate explanations that tie together many facts that would individually not be very interesting. A scientist who aimed only to accumulate truths and avoid errors would be awash in trivialities. Hence science aims for explanation as well as truth. These two goals subsume the goal of empirical adequacy, because for most scientists the point of describing and predicting observed phenomena is to find out what is true about them and to explain them.

But there are also practical goals that science accomplishes. Nineteenth-

century physicists such as Faraday and Maxwell were primarily driven by epistemic goals of understanding electrical and magnetic phenomena, but their work made possible the electronic technologies that now pervade human life. Research on topics such as superconductivity and lasers has operated with both scientific and technological aims. Molecular biology is also a field that began with primarily epistemic aims but that has increasingly been motivated by potential practical applications in medicine and agriculture. Similarly, the major focus of the cognitive sciences such as psychology and neuroscience has been understanding the basic mechanisms of thinking, but there have also been practical motivations such as improving education and the treatment of mental illnesses. It is clear, therefore, that one aim of the scientific enterprise is the improvement of human welfare through technological applications. This is not to say that each scientist must have that aim, since many scientists work far from areas of immediate application, but science as a whole has made and should continue to make technological contributions.

More critical views on the practical aims of science are extant. It has been claimed that science functions largely to help maintain the hegemony of dominant political and economic forces by providing ideologies and technologies that forestall the uprising of oppressed peoples. This claim is a gross exaggeration, but there is no question that the products of scientific research can have adverse effects—for example, the use of dubious theories of racial superiority to justify social policies and the use of advanced technology to produce devastating weapons. But saying that the aims of science are truth, explanation, and human welfare does not imply that these aims are always accomplished, only that these are the aims that science generally does and should have. We can now address the question of what strategies of rational thinking best serve the accomplishment of these aims.

Models of Individual Rationality

Consider a recent example of scientific reasoning, the collision theory of dinosaur extinction. Since the discovery of dinosaur fossils in the nineteenth century, scientists have pondered why the dinosaurs became extinct. Dozens of different explanations have been proposed, but in the past two decades one hypothesis has come to be widely accepted: dinosaurs became extinct around 65 million years ago because a large asteroid collided with the earth. Evidence for the collision hypothesis includes the discovery of a layer of iridium (a substance more common in asteroids than on earth) in geological formations laid down around the same
end p.365

time that the dinosaurs became extinct. What is the nature of the reasoning that led most paleontologists and geologists to accept the collision hypothesis and reject its competitors? I shall consider three main answers to this question, derived from confirmation theory, Bayesian probability theory, and the theory of explanatory coherence. In each case, I will describe a kind of ideal epistemic agent and consider whether scientists are in fact agents of the specified kind.

Confirmation and Falsification

Much work in the philosophy of science has presumed that scientists are *confirmation agents* that operate roughly as follows (see, e.g., Hempel 1965). Scientists start with hypotheses that they use to make predictions about observable phenomena. If experiments or other observations show that the predictions are true, then the hypotheses are said to be confirmed. A hypothesis that has received substantial empirical confirmation can be accepted as true, or at least as empirically adequate. For example, the hypothesis that dinosaurs became extinct because of an asteroid collision should be accepted if it as been confirmed by successful predictions.

Popper 1959 argues that scientists should not aim for confirmation but should operate as the following sort of *falsification agents*. Scientists use hypotheses to make predictions, but their primary aim should be to find evidence that contradicts the predicted results, leading to the rejection of hypotheses rather than their acceptance. Hypotheses that have survived severe attempts to falsify them are said to be corroborated. On this view, the proponents of the collision theory of dinosaur extinction should attempt to falsify their theory by stringent tests and only then consider them as corroborated, but not as accepted as true.

Although hypotheses are often used to make predictions, the process of science is much too complex for scientists to function generally as either confirmation agents or falsification agents. In particular, it is exceedingly rare for scientists to set out to refute their own hypotheses, and, given the difficulty of performing complex experiments, it is fortunate that they aim for confirmations rather than refutations. There are many reasons why an experimental prediction might fail, ranging from problems with instruments or

personnel to failure to control for key variables. A falsification agent would frequently end up throwing away good hypotheses.
But scientists are not just confirmation agents either, since hypotheses often get support not just from new predictions, but also from explaining data already obtained. Moreover, it often happens in science that there are conflicting hypotheses that are to some extent confirmed by empirical data. As Lakatos 1970 argues, the task then is not just to determine what hypotheses are confirmed, but also what hypotheses are better confirmed than their competitors. Hypothesis assessment is rarely a matter of evaluating a hypothesis with respect to its predictions, but rather requires evaluating competing hypotheses, with the best to be accepted and the others to be rejected. There are both probabilistic and explanatory approaches to such comparative assessment.

Probabilities

Carnap and numerous other philosophers of science have attempted to use the resources of probability theory to illuminate scientific reasoning (Carnap 1950 , Howson and Urbach 1989 , Maher 1993). Probabilistic agents operate as follows. They assess hypotheses by considering the probability of a hypothesis given the evidence, expressed as the conditional probability P(H/E). The standard tool for calculating such probabilities is Bayes's Theorem, one form of which is:

P(H/E) = P(H) * P(E/H) / P(E).

This says that the posterior probability of the hypothesis H given the evidence E is calculated by multiplying the prior probability of the hypothesis by the probability of the evidence given the hypothesis, all divided by the probability of the evidence. Intuitively, the theorem is very appealing, with a hypothesis becoming more probable to the extent that it makes improbable evidence more probable. Probabilistic agents look at all the relevant evidence, calculate values for P(E) and P(E/H), take into account some prior value of P(H), and then calculate P(H/E). Of two incompatible hypotheses, probabilistic agents prefer the one with the highest posterior probability. A probabilistic agent would accept the collision theory of dinosaur extinction if its probability given the evidence is higher than the probability of competing theories.
Unfortunately, it is not so easy as it sounds for a scientist to be a probabilistic agent. Various philosophers, such as Glymour (1980) and Earman (1992), have discussed technical problems with applying probability theory to scientific reasoning, but I will mention only what I consider to be the three biggest roadblocks. First, what is the interpretation of probability in P(H/E)? Probability has its clearest interpretation as frequencies in populations of observable events; for example, the probability that a die will turn up a 3 is 1/6, meaning that in a large number of trials there will tend to be 1 event in 6 that turns up a 3. But what meaning can we attach to the probability of dinosaur extinction being caused by an asteroid collision? There is no obvious way to interpret the probability of such causal hypotheses in terms of objective frequencies in specifiable populations.

The alternative interpretation is that such probabilities are degrees of belief,
end p.367

but there is substantial evidence that people's thinking does not conform to probability theory (see, e.g., Kahneman, Slovic, and Tversky 1982 ; Tversky 1994). One might say that the probability of a hypothesis is an idealized degree of belief, but it is not clear what this means. Degree of belief is sometimes cashed out in terms of betting behavior, but what would it mean to bet on the truth of various theories of dinosaur extinction?

The second difficulty in viewing scientists as probabilistic agents is that there are computational problems in calculating probabilities in accord with Bayes's theorem. In general, the problem of calculating probabilities is computationally intractable in the sense that the number of conditional probabilities required increases exponentially with the number of propositions. However, powerful and efficient algorithms have been developed for calculating probabilities in Bayesian networks that make simplifying assumptions about the mutual independence of different propositions (Pearl 1988). No one, however, has yet used Bayesian networks to simulate a complex case of scientific reasoning such as debates about dinosaur extinction. In contrast, the next section discusses a computationally feasible account of scientific inference based on explanatory coherence.

The third difficulty with probabilistic agents is that they may ignore qualitative factors affecting theory choice. Scientists' arguments suggest that they care not only how much evidence there is for a theory, but also about the variety of the evidence, the simplicity of the theory that accounts for it, and analogies between proposed explanations and other established ones. Perhaps simplicity and analogy could be accounted for in terms of prior probabilities: a simpler theory or one offering analogous explanations would get a higher value for P(H) to be fed into the calculation via Bayes's theorem of the posterior probability P(E/H). But the view of probability as subjective degree of belief leaves it mysterious how people do or should arrive at prior probabilities.

Explanatory Coherence

If scientists are not confirmation, falsification, or probabilistic agents, what are they? One answer, which goes back to two nineteenth-century philosophers of science, William Whewell and Charles Peirce, is that they are *explanation agents.* On this view, what scientists do in theoretical inference is to generate explanations of observed phenomena, and a theory is to be preferred to its competitors if it provides a better explanation of the evidence. Theories are accepted on the basis of an *inference to the best explanation.* Such inferences are not merely a matter of counting which among competing theories explains more pieces of evidence, but also require assessment in terms of the overall explanatory coherence of each
end p.368

hypothesis with respect to a scientist's whole belief system. Factors that go into this assessment for a particular hypothesis include the evidence that it explains, its explanation by higher-level hypotheses, its consistency with background information, its simplicity, and analogies between the explanations offered by the hypothesis and explanations offered by established explanations (Harman 1986 , Lipton 1991 , Thagard 1988).

The major difficulty with the conception of scientists as explanatory agents is the vagueness of concepts such as explanation, inference to the best explanation, and explanatory coherence. Historically, explanation has been conceptualized as a deductive relation, a probabilistic relation, and a causal relation. The deductive conceptualization of explanation fits well with the confirmation and falsification view of agents: a hypothesis explains a piece of evidence if a description of the evidence follows deductively from the hypothesis. Similarly, the probabilistic conceptualization of explanation fits well with the probabilistic view of agents: a hypothesis explains a piece of evidence if the probability of the evidence given the hypothesis is higher than the probability of the evidence without the hypothesis. Like Salmon (1984) and others, I prefer a conceptualization of explanation as the provision of causes: a hypothesis explains a piece of evidence if it provides a cause of what the evidence describes. The causal conceptualization must face the problem of saying what causes are and how causal relations are distinct from deductive and probabilistic ones (see Thagard 1999 , chap. 7).

Assuming we know what an explanation is, how can we characterize inference to the best explanation? I have shown how a precise and easily computable notion of explanatory coherence can be applied to many central cases in the history of science (Thagard 1992). For example, we can understand why the collision theory of dinosaur extinction has been accepted by many scientists but rejected by others by assessing its explanatory coherence with respect to the evidence available to different scientists (see Thagard 1991 for computer simulations of the dinosaur debate using the program ECHO).

I prefer to view scientists as explanation agents rather than as confirmation, falsification, or probabilistic agents because this view fits better with the historical practice of scientists as evident in their writings, as well as with psychological theories that are skeptical about the applicability of deductive and probabilistic reasoning in human thinking. But I acknowledge that the probabilistic agent view is probably the most popular one in contemporary philosophy of science; it has largely absorbed the confirmation agent view by the plausible principle that evidence confirms a hypothesis if and only if the evidence makes the hypothesis more probable—that is, $P(H/E) > P(H)$. It is also possible that scientists are not rational agents of any of these types but rather are reasoners of a very different sort. For example, Mayo 1996 develops a view of scientists as modeling patterns of experimental results that are useful for distinguishing errors.

Solomon 2001

end p.369

describes scientists as reaching conclusions based on a wide variety of “decision vectors,” ranging from empirical factors such as salience of data to nonempirical factors such as ideology.

Practical Reason

As mentioned in this chapter's introduction, there is much more to scientific rationality than accepting and rejecting hypotheses. Here are some of the important decisions that scientists make in the course of their careers:

1. What general field of study should I enter—for example, should I become a paleontologist or a geologist?
2. Where and with whom should I study?
3. What research topics should I pursue?
4. What experiments should I do?
5. With whom should I collaborate?

When scientists make these decisions, they are obviously acting for more than epistemic reasons, entering a field for more reasons than that it would maximize their stock of truths and explanations. Scientists have personal aims as well as epistemic ones, such as having fun, being successful, living well, becoming famous, and so on. Let us now consider two models of scientists as practical decision makers: scientists as utility agents and scientists as emotional agents.

The utility agent view is the familiar one from economics, with an agent performing an action because of a calculation that the action has more expected utility than alternative actions, where expected utility is a function of the utilities and probabilities of different outcomes. This view is consonant with the epistemic view of scientists as probabilistic agents and has many of the same difficulties. When scientists are considering between different research topics, do they have any idea of the relevant probabilities and utilities? Suppose I am a molecular biologist doing genome research and have to decide whether to work with yeast or with worms. I may have hunches about which research program may yield the more interesting results, but it is hard to see how these hunches could be translated into anything as precise as probabilities and utilities.

A more realistic view of the decision making of scientists and people in general is that we choose the actions that receive the most positive emotional evaluation based on their coherence with our goals (Thagard 2000 , chap. 6; Thagard 2001). On this view, decision making is based on intuition rather than on numerical calculation: unconsciously we balance different actions and different goals, arriv

end p.370

ing at a somewhat coherent set of accepted ones. The importance of goals is affected by how they fit with other goals as well as with the different actions that are available to us. We may have little conscious awareness of this balancing process, but the results of the process come to consciousness via emotions. For example, scientists may feel excited by a particular research program and bored or even disgusted by an alternative program. Psychologists use the term *valence* to refer to positive or negative emotional evaluations. For discussions of the role of emotions in scientific thinking, see Thagard (2002a , b). Like Nussbaum (2001), I view emotions as intelligent reactions to perceptions of value, including epistemic value.

Just as there is a concordance between the probabilistic view of epistemic agents and the utility view of practical agents, there is a concordance between the explanatory coherence view of epistemic agents and the emotional coherence view of practical agents. In fact, emotions play a significant role in inference to hypotheses as well as in inference to actions, because the inputs to and outputs from both kinds of inference are emotional as well as cognitive. The similarity of outputs is evident when scientists appreciate the great explanatory power of a theory and characterize it as elegant, exciting, or even beautiful. As with practical judgments of emotional coherence in practical decision making, we have no direct conscious access to the cognitive processes by which we judge some hypotheses to be more coherent than others. What emerges to consciousness from a judgment of explanatory coherence is often emotional, in the form of liking or even joy with respect to one hypothesis, and dislike or even contempt for rejected competing hypotheses. For example, when Walter and Luis Alvarez came up with the hypothesis that dinosaurs had become extinct because of an asteroid collision, they found the hypothesis not only plausible but even exciting (Alvarez 1998). In contrast, some skeptical paleontologists thought the hypothesis was not only dubious but also ridiculous. Emotional inputs to hypothesis evaluation include the varying attitudes that scientists hold toward different experimental results and even for different experiments—any good scientist knows some experiments are better than others. Another kind of emotional input is analogical: a theory analogous to a positively viewed theory such as evolution will have greater positive valence than one that is analogous to a scorned theory such as cold fusion.

Thus my view of scientists as explanatory-emotional agents is very different from the view of them as probabilistic-utility agents. My emphasis on emotions will probably have readers wondering whether scientists are rational at all. Perhaps they are just swayed by their various intellectual prejudices and personal desires to plan research programs and accept hypotheses in ways that disregard the epistemic aims of truth and explanation. There are, unfortunately, cases where scientists are deviant in these ways, with disastrous results such as fraud and other kinds of unethical behavior. But the temperaments and training of most scientists are such that they have an emotional attachment to the crucial epistemic aims. Many scientists become scientists because they enjoy finding out how things work, so that the aims of truth and explanation are with them from the beginnings of their scientific training. These attachments can be fostered by working with advisors who not only value these aims but also transmit their emotional evaluations to the students and postdoctoral fellows with whom they work. So, for most scientists, a commitment to fostering explanation and truth is an emotional input into their practical decision making.

Models of Group Rationality

As Kuhn (1970) and many other historians, philosophers, and sociologists of science have noted, science is not merely a matter of individual rationality. Scientists do their work in the context of groups of various sizes, from the research teams in their own laboratories to the community of scientists working on similar projects to the overall scientific community. As I have documented elsewhere (Thagard 1999 , chap. 11), most scientific articles have multiple authors, and the trend is toward increasing collaboration. In addition, all scientists operate within the context of a wider community with shared societies, journals, and conferences. Therefore the question of the rationality of science can be raised for groups as well as individuals: What is it for a group of scientists to be collectively rational, and are such groups generally rational? I will assume that groups of scientists have the same primary aims that I attributed to science in general: truth, explanation, and human welfare via technological applications.

It might seem that the rationality of scientific groups is just the sum of the rationality of the individuals that comprise them. Then a group is rational if and only if the individual scientists in it are rational. But it is possible to have individual rationality without group rationality, if the pursuit of scientific aims by each scientist does not add up to optimal group performance. For example, suppose that each scientist rationally chooses to pursue exactly the same research strategy as the others, with the result that there is little diversity in the resulting investigations, and paths that would be more fertile with respect to truth and explanation are not taken. Philosophers such as Kitcher (1993) have emphasized the need for cognitive diversity in science.

On the other hand, it might be possible to have group rationality despite lack of individual rationality. Hull (1989) has suggested that individual scientists who seek fame and power rather than truth and explanation may in fact contribute to the overall aims of science, because their individualistic pursuit of nonepistemic

end p.372

motives in fact leads the scientific group as a whole to prosper. This is analogous to Adam Smith's economic model in which individual greed leads to overall economic growth and efficiency.

It is important to recognize also that group rationality in science is both epistemic and practical. Of a particular scientific community, we can ask two kinds of question:

(1)Epistemic: Given the evidence, what should be the distribution of beliefs in the community?
(2)Practical: What should be the distribution of research initiatives in the community?

For the epistemic question, it might be argued that if all scientists have access to the same evidence and hypotheses, then they should all acquire the same beliefs. Such unanimity would, however, be detrimental to the long-term success of science, since it would reduce cognitive diversity. For example, if Copernicus had been enmeshed within the Ptolemaic theory of the universe, he might never have generated his alternative heliocentric theory, which turned out to be superior with respect to both truth and explanation. Similarly, in the dinosaur case Walter Alvarez would never have formulated his theory of why dinosaurs became extinct if he had been a conventional paleontologist.

Moreover, epistemic uniformity would contribute to practical uniformity, which would clearly be disastrous. It would be folly to have all scientists within a scientific community following just a few promising leads, since this would reduce the total accomplishment of explanations as well as retard the development of novel explanations. Garrett Hardin (1968) coined the term "tragedy of the commons" to describe a situation in which individual rationality could promote group irrationality. Consider sheep herders who share a common grazing area. Each herder separately may reason that adding one more sheep to his or her herd would not have any serious effect on the common area. But such individual decisions might collectively produce overgrazing, so that there is not enough food for any of the sheep, with the result that all sheep herders are worse off. Analogously, we can imagine in science and other organizations a kind of "tragedy of consensus," in which the individuals all reach similar conclusions about what to believe, stifling creative growth.

So, what should be our model of group rationality in science? Kitcher 1993 and Goldman 1999 develop models of group rationality that assume that individual scientists are probabilistic agents. Although these analyses are interesting with respect to cognitive diversity and truth attainment, I do not find them plausible because of the problems with the probabilistic view discussed in the last section. As an alternative, I have developed a model of scientific consensus based on explanatory coherence.

This model is called CCC, for *consensus* = *coherence* + *communication* (Tha
end p.373

gard 2000 , chap. 10). It assumes that each scientist is an explanation agent, accepting and rejecting hypotheses on the basis of their explanatory coherence with evidence and alternative hypotheses. Communication takes place as the result of meetings between scientists in which they exchange information about available evidence and hypotheses. If all scientists acquire exactly the same information, then they will agree about what hypotheses to accept and reject. However, in any scientific community, exchange of information is not perfect, so that some scientists may not hear about some of the evidence and hypotheses. Moreover, different scientists have different antecedent belief systems, so the overall coherence of a new hypothesis may be different for different scientists. Ideally, however, if communication continues there will eventually be community consensus as scientists accumulate the same sets of evidence and hypotheses and therefore reach the same coherence judgments. The CCC model has been implemented as a computational extension of the explanatory coherence program ECHO, in which individual scientists evaluate hypotheses on the basis of their explanatory coherence but also exchange hypotheses and evidence with other scientists. These

simulated meetings can be either pairwise exchanges between randomly selected pairs of scientists or "lectures" of the sort that take place at scientific conferences in which one scientist can broadcast sets of hypotheses and evidence to a group of scientists. Of course, communication is never perfect, so it can take many meetings before all scientists acquire approximately the same hypotheses and evidence. I have performed computational experiments in which different numbers of simulated scientists with varying communication rates achieve consensus in two interesting historical cases: theories of the causes of ulcers, and theories of the origins of the moon.

The CCC model shows how epistemic group rationality can arise in explanation agents who communicate with each other, but it tells us nothing about practical group rationality in science. One possibility would be to attempt to extend the probabilistic-utility model of individual practical reason. On this model, each scientist makes practical decisions about research strategy based on calculations concerning the expected utility of different courses of action. Research diversity arises because different scientists attach different utilities to various experimental and theoretical projects. For reasons already given, I would prefer to extend the explanatory-emotional model described in the previous section.

The extension arises naturally from the CCC model just described, except that in large, diverse communities we should not expect the same degree of practical consensus as there is of epistemic consensus, for reasons given below. For the moment, let us focus on particular research groups rather than on whole scientific communities. At this level, we can find a kind of local consensus that arises because of emotional coherence and communication. The characteristics of the group include the following:

end p.374

1. Each scientist is an explanation agent with evidence, hypotheses, and the ability to accept and reject them on the basis of explanatory coherence.
2. In addition, each scientist is an emotional agent with actions, goals, valences, and the ability to make decisions on the basis of emotional coherence.
3. Each scientist can communicate evidence and hypotheses with other scientists.
4. Each scientist can, at least sometimes, communicate actions, goals, and valences to other scientists.
5. As the result of cognitive and emotional communication, consensus is sometimes reached about what to believe and also about what to do.

The hard part to implement is the component of (4) that involves valences. It is easy to extend the CCC model of consensus to include emotional coherence simply by allowing actions, goals, and valences to be exchanged just like evidence, hypotheses, and explanations.

In real life, valences are not so easily exchanged as verbal information about actions, goals, and what actions accomplish which goals. Just hearing someone say that they really care about something does not suffice to make you care about it too, nor should it, because your goals and valences may be orthogonal or even antagonistic to mine. So in a

computational model of emotional consensus, the likelihood of exchange of goals and valences in any meeting should be much lower than the likelihood of exchange of hypotheses, evidence, and actions.

Still, in real-life decision making involving scientists and other groups such as corporate executives, emotional consensus is sometimes reached. What are the mechanisms of valence exchange—that is, how do people pass their emotional values on to other people? Two relevant social mechanisms are emotional contagion and attachment-based learning. Emotional contagion occurs when person A expresses an emotion and person B unconsciously mimics A's facial and bodily expressions and then begins to acquire the same emotion (Hatfield, Cacioppo, and Rapson 1994). For example, if a group member enthusiastically presents a research strategy, then the enthusiasm may be conveyed through both cognitive and emotional means to other members of the group. The cognitive part is that the other group members become aware of possible actions and their potential good consequences, and the emotional part is conveyed by the facial expressions and gestures of the enthusiast, so that the positive valence felt by one person spreads to the whole group. Negative valence can also spread, not just from a critic pointing out drawbacks to a proposed action as well as more promising alternatives, but also by contagion of the negative facial and bodily expressions.

Another social mechanism for valence exchange is what Minsky (2001) calls attachment-based learning. Minsky points out that cognitive science has developed good theories of how people use goals to generate subgoals, but has had little to

say about how people acquire their basic goals. Similarly, economists employing the expected-utility model of decision making take preferences as given, just as many philosophers who hold a belief-desire model of rationality take desires as given. Minsky suggests that basic goals arise in children as the result of praise from people to whom the children are emotionally attached. For example, when young children share their toys with their playmates, they often receive praise from their parents or other caregivers. The parents have positive valence for the act of sharing, and the children may also acquire a positive emotional attitude toward sharing as the result of seeing that it is something cared about by people whom they care about and who care about them. It is not just that sharing becomes a subgoal to accomplish the goal of getting praised by parents; rather, being kind to playmates becomes an internalized goal that has intrinsic emotional value to the children.

I conjecture that attachment-based learning also occurs in science and other contexts of group decision making. If your supervisor is not just a boss but also a mentor, then you may form an emotional attachment that makes you particularly responsive to what the supervisor praises and criticizes. This makes possible the attachment-based transmission of positive values such as zeal for truth and understanding, or, more locally, for integrity in dealing with data and explanations.

Notice that both emotional contagion and attachment-based learning require quite intense interpersonal contacts that will not be achieved in a large lecture hall or video conference room, let alone through reading a published article. The distinguished social psychologist, Richard Nisbett, told me that he learned how to do good experiments through discussions

with his supervisor, Stanley Schacter. Nisbett said, “He let me know how good my idea was by grunts: noncommittal (`hmmm'), clearly disapproving (`ahnn') or (very rarely) approving (`ah!').” These grunts and their attendant facial expressions conveyed emotional information that shaped the valences of the budding researcher.
Accordingly, when I extend my CCC model of consensus as coherence plus communication to include group decisions, I will include two new variables to determine the degree of valence transmission between agents: degree of personal contact, and degree of attachment. If personal contact and attachment are high, then the likelihood of valence transmission will be much greater than in the ordinary case of scientific communication, in which the success of verbal transmission of information of hypotheses, evidence, and actions is much higher than the transmission of valences.
There may, however, be quasi-verbal mechanisms for valence transfer. Thagard and Shelley (2001) discuss emotional analogies whose purpose is to transfer valences as well as verbal information. For example, if a scientist presents a research project as analogous to a scientific triumph such as the asteroid theory of dinosaur extinction, then listeners may transfer the positive value they feel for the asteroid theory to the proposed research project. Alternatively, if a project is analogous to the cold fusion debacle, then the negative valence attached to that case may be projected onto the proposed project. Thus emotional analogies are a third mechanism, in addition to emotional contagion and attachment-based learning, for transfer of valences. All three mechanisms may interact with each other, for example when a mentor uses an emotional analogy and facial expressions to convey values to a protégé. Alternatively, the mentor may function as a role model, providing a different kind of emotional analogy: students who see themselves as analogous to their role models may tend to transfer to themselves some of the motivational and emotional characteristics of their models.
I hope it is obvious from my discussion of practical group rationality in science why science need not succumb to the tragedy of consensus, especially with respect to practical rationality. Communication between scientists is imperfect, both with respect to cognitive information such as hypotheses and evidence and especially with respect to emotional valences for particular approaches. Scientists may get together for consensus conferences such as the ones sponsored by the National Institutes of Health that regularly deal with controversial issues in medical treatment (see Thagard 1999 , chap. 12, for a discussion). But not all scientists in a community attend such conferences or read the publications that emanate from them. Moreover, the kinds of close interpersonal contact needed for communication of values by emotional contagion and attachment-based learning occur only in small subsets of the whole scientific community. Hence accomplishment of the general scientific aims of truth, explanation, and technological applications need not be hindered in a scientific community by a dearth of practical diversity. Solomon 2001 provides a rich discussion of consensus and dissent in science.

Is Science Rational?

A person or group is rational to the extent that its practices enable it to accomplish its legitimate goals. At the beginning of this chapter, I argued that the legitimate goals of

science are truth, explanation, and technologies that promote human welfare. Do scientific individuals and groups function in ways that further these goals, or do they actually pursue other personal and social aims that are orthogonal or even antagonistic to the legitimate goals? I will now consider several psychological and sociological challenges to the rationality of science.

Psychological challenges can be based on either cold cognition, which involves processes such as problem solving and reasoning, or hot cognition, which includes

end p.377

emotional factors such as motivation. The cold-cognition challenge to scientific rationality would be that people's cognitive processes are such that it is difficult or impossible for them to reason in ways that promote the aims of science. If scientific rationality required people to be falsification agents or probabilistic agents, then the cold-cognition challenge would be a serious threat. I cited earlier some of the experimental and historical data that suggest that probabilistic reasoning and falsification are not natural aspects of human thinking. In contrast, there is evidence that people can use explanatory coherence successfully in social judgments (see Read and Marcus-Newhall 1993).

One might argue that there is evidence that people are confirmation agents, and not very good ones in that they tend toward *confirmation bias* in looking excessively to confirm their hypotheses rather than falsify them (see Klayman and Ha 1987). However, the psychological experiments that find confirmation biases involve reasoning tasks that are much simpler than those performed by actual scientists. Typically, nonscientific subjects are asked to form generalizations from observable data, for example in seeing patterns in numerical sequences. The generalization tasks of real scientists are more complex, in that data interpretation requires determining whether apparent patterns in the data are real or just artifacts of the experimental design. If scientists did not try hard to get their experiments to confirm their hypotheses, the experiments would rarely turn out to be interesting. Notably, trying hard to confirm is not always sufficient to produce confirming results, so scientists sometimes have falsification thrust upon them. But their bias toward finding confirmations is not inherently destructive to scientific rationality.

A more serious challenge to the rationality of science comes from hot cognition. Like all people, scientists are emotional beings, and their emotions may lead to distortions in their scientific works if they are attached to values that are inimical to the legitimate aims of science. Here are some kinds of cases where emotions have distorted scientific practice:

1. Scientists sometimes advance their own careers by fabricating or distorting data in order to support their own hypotheses. In such cases, they have greater motivation to enhance their own careers than to pursue truth, explanation, or welfare.
2. Scientists sometimes block the publication of theories that challenge their own by fabricating problems with submitted articles or grant proposals that they have been asked to review.
3. Without being fraudulent or intentionally evil, scientists sometimes unintentionally deceive themselves into thinking that their hypotheses and data are better than those of

1. Scientists sometimes advance their own careers by fabricating or distorting data in order to support their own hypotheses. In such cases, they have greater motivation to enhance their own careers than to pursue truth, explanation, or welfare.
   their rivals.
4. Scientists sometimes further their careers by going along with politically mandated views—for example, the Nazi rejection of Einsteinian physics and the Soviet advocacy of Lysenko's genetic theories.

end p.378

Cases like these show indubitably that science is not always rational. Some sociologists such as Latour (1987) have depicted scientists as largely concerned with gaining power through the mobilization of allies and resources.

It is important to recognize, however, that the natural emotionality of scientists is not in itself a cause of irrationality. As I documented elsewhere, scientists are often motivated by emotions that further the goals of science, such as curiosity, the joy of discovery, and appreciation of the beauty of highly coherent theories (Thagard, 2002b). Given the modest incentive structure of science, a passion for finding things out is a much more powerful motivator of the intense work required for scientific success than are extrinsic rewards such as money and fame. Thus hot cognition can promote scientific rationality, not just deviations from it. The mobilization of resources and allies can be in the direct or indirect service of the aims of science, not just the personal aims of individual scientists.

A useful response to the question "Is science rational?" is "Compared to what?" Are scientists as individuals more adept than nonscientists at fostering truth, explanation, and human welfare? The history of science and technology over the past two hundred years strongly suggests that the answer is yes. We have acquired very broadly explanatory theories such as electromagnetism, relativity, quantum theory, evolution, germ theory, and genetics. Thousands of scientific journals constitute an astonishing accumulation of truths that ordinary life would never have allowed. Moreover, technologies such as electronics and pharmaceuticals have enriched and lengthened human lives. So the occasional irrationality of individual scientists and groups is compatible with an overall judgment that science is in general a highly rational enterprise.

In recent decades, the most aggressive challenge to the ideal of scientists as rational agents has come from sociologists and historians who claim that scientific knowledge is "socially constructed." Obviously, the development of scientific knowledge is a social as well as an individual process, but the social construction thesis is usually intended to make the much stronger claim that truth and rationality have nothing to do with the development of science. My own view is that an integrated psychological/sociological view of the development of scientific knowledge is perfectly compatible with scientific rationality involving the frequently successful pursuit of truth, explanation, and human welfare (Thagard 1999).

Crucially, however, the assessment of scientific rationality needs to employ models of individual reasoning and group practices that reflect the thought processes and methodologies of real scientists. Models based on formal logic and probability theory have tended to be so remote from scientific practice that they encourage the inference that

scientists are irrational. In contrast, psychologically realistic models based on explanatory and emotional coherence, along with socially realistic models of consensus, can help to illuminate the often impressive rationality of the enterprise of science.
end p.379

chapter 20 ECONOMIC RATIONALITY

Paul. Weirich

According to an introductory economics textbook by Case and Fair (1992 , 5), "Economics is the study of how individuals and societies choose to use the scarce resources that nature and previous generations have provided." Economic theory assumes as a first approximation that individuals and societies choose ration-ally. Consequently, it carefully formulates principles of rational decision making.

I examine conceptions of rationality common in economics. Although economists treat rationality from multiple perspectives, I describe only a few central views. Other handbook entries supplement my sketch. The title "Economic Rationality" is not meant to suggest that rationality comes in various types, one of which is appropriate in economics. The familiar concept of rationality, shared by all the disciplines, guides principles of rationality in economics.

Economics divides into micro- and macroeconomics. The former treats the interaction of consumers and producers, such as price setting, and the latter treats the aggregate results of that interaction, such as the national inflation rate. Microeconomics formulates principles of rational choice. It studies both individual and group decision making. The division between these types of decision is flexible, however. A household or a firm composed of individuals may be treated as a single individual.

Economists focus on rationality in choice and action. Rationality in preference and belief arise as ancillary topics. They treat instrumental rationality—that is, the rational adoption of means to reach ends. They also propose standards of
end p.380

rationality demanding various types of consistency. For example, Raiffa (1968 , 76), Pindyck and Rubinfeld (1989 , 59), and Case and Fair (1992 , 184) assume that rationality requires the consistency of a preference ranking, understood as the transitivity of preferences and of attitudes of indifference. Consistency of this type goes beyond logical consistency.

A household has diverse goals. Besides food, clothing, and shelter, it may want education and entertainment. A firm typically has more limited goals such as profit and reduction of risk. Economists tend to leave to philosophers the formulation of standards of rationality for basic goals, but they assume that the usual goals of households and firms are rational. Economists may seem to condemn as irrational certain basic goals. They commonly warn against the irrationality of being swayed by sunk costs, neglecting opportunity costs, and having pure time preferences. A textbook may, for example, admonish the owner of a languishing restaurant for sinking more money into his business, trying to keep it afloat

against the odds, instead of closing it and accepting the loss of his initial investment. It may criticize a married couple who continue to occupy a large house after their children have grown up and moved away, ignoring the opportunity to sell their house and use the proceeds to buy both another, smaller house and a retirement annuity. It may reprimand an impatient person who picks the lesser of two goods just because it is available a little earlier. One's impression may be that the proscribed behavior is deemed irrational because it proceeds from irrational basic goals. For instance, it may seem that counting sunk costs is deemed irrational because it proceeds from an irrational basic goal never to accept a loss. However, economists typically hold that such mistakes stem from misperception of options and consequences, or muddled thinking. They count the mistakes as errors of judgment rather than as errors in basic goals and so do not regard their warnings as criticisms of basic goals.

1. Conceptions of Rationality in Economics

Three conceptions of rationality dominate recent economic theorizing. They are closely related and may be thought of as links in an evolutionary chain rather than as rivals for the same environmental niche. Maximizing Self-Interest

A traditional view in economics takes a person's basic goal to be self-interest and takes rationality to be the promotion of self-interest. When several courses of action are open, rationality recommends the one that best promotes self-interest. It says to maximize self-interest. When the action that maximizes self-interest is uncertain, rationality recommends probability as a guide. Maximize expected self-interest, it says.

Edgeworth clearly enunciated the assumption of self-interest. According to him, “The first principle of economics is that every agent is actuated only by self-interest” (1881 , 16). Philosophers call this view psychological egoism. A closely related view is egoism about rationality, the view that rationality requires a person to promote her own interest exclusively.

Some presentations of the egoistic view of rationality use technical terminology. Instead of speaking of an agent's interest, they speak of utility for her. Pindyck and Rubinfeld (1989 , 81, 87) define utility for an agent as satisfaction for her and recommend that an agent maximize utility for herself, that is, maximize her satisfaction. Their analyses assume that a firm seeks long-run profits and that rationality requires its acting to maximize long-run profits (246–47, 458). Case and Fair (1992 , 88, 160, 181, 476–77) take utility as happiness, satisfaction, or well-being and say that an agent maximizes, and ought to maximize, utility for herself. Although psychological egoism and egoism are common in economics, they may be adopted as simplifying assumptions met approximately in typical economic transactions.

Economists point out that, contrary to common belief, self-interest does not generate a war of all against all. One of the lessons of Adam Smith's *The Wealth of Nations* is that agents pursuing self-interest are often led to cooperate because cooperation is frequently an excellent means of promoting their own interests. Self-interested agents may trade goods, for instance, because each benefits from the exchange (Smith 1976 , 26–27). Smith proclaims that self-interest leads agents to act as if an invisible hand were guiding

their behavior toward the common good (1976 , 456). In the same vein, as Pindyck and Rubinfeld (1989 , 570) explain, contemporary welfare economics shows that if everyone trades in the marketplace to maximize her satisfaction, and all mutually beneficial trades are completed, the resulting allocation will be efficient in the sense that no alternative allocation yields gains for some without losses for others. Efficiency in this sense is called Pareto optimality after the nineteenth-century Italian economist Vilfredo Pareto.
Self-interest has different meanings for different theorists. The standard interpretation takes a person's interest as her welfare or well-being, that is, what is good for her. However, those holding that all people are exclusively self-interested sometimes take a person's interest as satisfaction of her desires, whatever their content. Fulfilling any of her desires counts as promoting her interest. Then the
end p.382

claim that people are self-interested says no more than that people pursue their desires. As Sen (1977 , 322–24) observes, it does not rule out altruism, since an altruist pursues her desire to help others.
Binmore (1994 , 6, 15, 18–19, 21) claims that rationality amounts to promoting enlightened self-interest, or broad self-interest. He means promotion of one's goals whatever they may be, even if they include showing sympathy toward others and honoring commitments (21, 28). Taking self-interest broadly makes promotion of self-interest generate the next conception of rationality.

Maximizing Utility

During the twentieth century, economists discovered techniques for using an agent's preferences among acts to derive his assignment of probabilities and utilities to the acts' possible outcomes and thereby derive the acts' utilities. Von Neumann and Morgenstern (1953) developed and promulgated such methods, but Ramsey (1931 , chap. 7) had worked out similar methods earlier (see Joyce, chap. 8, and Dreier, chap. 9, this volume). The methods derive subjective probabilities and utilities—roughly, degrees of belief and degrees of desire, respectively. They enrich accounts of decision making in the face of risk. They also provide a framework for a broad account of an agent's basic goals. The new utility theory does not restrict the content of those goals. According to Ramsey, “The theory I propose to adopt is that we seek things which we want, which may be our own or other people's pleasure, or anything else whatever, and our actions are such as we think most likely to realize these goods” (1931, 173). Modern utility theory generalizes maximization of self-interest to maximization of personal utility. It takes utility broadly enough to accommodate all goals, not just self-interest.
Von Neumann and Morgenstern (1953 , 1, 8–9) say economic tradition characterizes rational behavior as the maximization of utility taken as satisfaction in the case of the consumer and profit in the case of the entrepreneur. However, the utility theory they articulate (chap. 3) does not restrict an agent's goals. Savage (1972), who extended von Neumann and Morgenstern's approach to utility, says that the kind of utility they

introduce is dependent on probability in contrast with earlier kinds—for example, satisfaction—that are independent of probability. Their probability-dependent type of utility figures in the representation of preferences among acts involving risk and, he says, is better able to explain rational behavior (1972, 91–104). It does not assume that an agent has selfish goals; it just arithmetizes the relation of preference among acts (1972, 69). Von Neumann and Morgenstern make utility dependent on objective probability, whereas, for greater explanatory power, Savage makes utility dependent on subjective probability.
end p.383

Although as Arrow (1963 , 3) observes, economic tradition takes rationality to be some sort of maximization, modern utility theory in taking rationality as maximization of utility departs from the view that rationality is maximization of self-interest. An agent's utility assignment need not follow self-interest. Other goals, including concern for other people, may influence an agent's utility assignment. According to the new view, rationality recognizes all of an agent's goals. Harsanyi (1977 , 10), working within the same school, recognizes a utility function that assigns positive utility to unselfish values. He says that rational agents pursue both the selfish and unselfish values to which their utility functions assign positive utility.

Sometimes in modern utility theory maximizing utility goes by other names. Maximizing utility among risky options may be called maximizing expected utility since it depends on probabilities and expectations concerning options' possible outcomes. Maximizing utility and maximizing expected utility are equivalent, however. Utility, being subjective, is relative to information. Because an agent's utility assignment to a risky option depends on the agent's goals and information, changes in the agent's information influence the option's utility assignment. The more evidence the option will realize the agent's goals, the more attractive it is to him. An option's utility and expected utility therefore agree.

Sometimes utility is taken to be relative to full information. Under this interpretation, utility may differ from expected utility. Expected utility is relative to actual information, not necessarily full information. Since utility relative to full information may differ from utility relative to actual information, it may differ from expected utility even though utility relative to actual information agrees with expected utility. I treat maximizing utility with respect to actual information, which is the same as maximizing expected utility, and so do not distinguish between maximizing utility and maximizing expected utility.

Because utility represents preference, maximizing utility among options is the same as adopting an option at the top of one's preference ranking of options. Thus, Varian's (1984 , 115) view that rationality requires preference maximization agrees with the view that it requires utility maximization. The difference between preference maximization and utility maximization is only terminological.

Because modern utility theory derives utilities from preferences, it often bills utility maximization as a form of consistency among preferences, or among choices, or among preferences and choices. Luce and Raiffa (1957 , 31–32) say that it takes utility as a representation of preferences or a guide to consistent action. Barry (1965 , 4–5) says that economists hold that rationality does not require maximizing utility, thought of as

satisfaction, but only requires a consistent pattern of choices. Arrow (1967 , 5) says that rationality is a matter of making choices derivable from a preference ordering. This condition imposes a type of consistency on choices. Harsanyi (1977 , 8) also says that choice behavior is rational if it satisfies certain consistency requirements. According to Broome (1991 , 90–92), modern utility theory claims only that a preference ranking, if rational, may be represented
end p.384

by a utility assignment according to which an act's utility equals the expected utility of its possible outcomes. This condition imposes a type of consistency on a preference ranking. Binmore (1994 , 21, 27) says that rationality requires only being consistent in seeking one's goals, which demands consistency between preferences and acts.

The train of thought that construes utility maximization as a form of consistency is roughly the following. The techniques for deriving probabilities and utilities from preference rankings assume that preference rankings are transitive and meet other requirements that may be taken as requirements of consistency in a broad sense. In general, the techniques require preference rankings to be *as if* maximizing utility; preferences among options have to follow their utilities computed with respect to the derived probability and utility assignments for the options' possible outcomes. This global requirement may also be taken as a consistency requirement. It yields utility maximization given utility's derivation from preference rankings.

Operationists go further. They define probability and utility in terms of preferences, and some, such as Binmore (1994 , 169), moreover define preferences in terms of choices. As Binmore notes, since operationist definitions of preferences and utilities make them deducible from choices, satisfying preferences or maximizing utility is just consistent choice behavior (1994 , 50–51). For example, consider an operational definition of utility as a mathematical representation of preferences according to which preferences follow utilities. Because the definition entails that preferences follow utilities, it makes that traditional consistency requirement a definitional truth and deprives it of normative force. Similarly, a definition of preference in terms of choice makes the traditional consistency requirement that choice follow preference a definitional truth without normative force. That choices maximize utility becomes a definitional truth. After operationalization, the only consistency requirements with normative force concern consistency of preferences or consistency of choices. See Blackburn 1998 , chap. 6, and Hubin 2001 , sec. 3.

Bounded Rationality

The previous sections examined conceptions of rationality requiring maximization of some sort. However, some theorists doubt that humans are capable of the kind of maximization demanded. Simon (1982) advocates standards of rationality attuned to human limits, especially cognitive limits. His main proposal starts with the observation that people do not survey all their options at the start of a decision problem. They search

for options during deliberations and entertain them sequentially. Also, to simplify evaluation of an option they just classify it as satis
end p.385

factory or not satisfactory, depending on whether it meets their aspirations. To approximate optimization in a practical way, Simon proposes adopting the first satisfactory option discovered (1982, 250–51). Someone selling a house, for example, may rationally accept the first satisfactory offer instead of holding out for an optimal offer. Someone playing chess may adopt the first satisfactory move spotted instead of attempting the Herculean task of identifying the optimal move. Simon calls his procedure "satisficing" and calls theories incorporating such procedures theories of "bounded rationality."

Simon argues that in view of our limited time and cognitive capacity for making decisions, we should not be held to the classical standard of utility maximization. In general, he advocates replacing substantive rules of rationality, such as the rule to maximize utility, with procedural rules of rationality, such as the rule to satisfice (1982, 424–43). Procedural rules prescribe methods of making decisions rather than decisions to be made. They focus on the process rather than the result. They consider how decisions should be made rather than what decisions should be made. Some commonsense rules for making decisions are procedural. The advice to sleep on an important decision before finalizing it, for instance, attends to method rather than result. According to Simon, whereas substantive rationality targets best solutions, procedural rationality targets solutions good in view of human limits. Procedural rationality is practical in the face of problems that are intractable from the perspective of substantive rationality (1982, 428, 431). [1]

The rule to satisfice is not purely procedural, since it aims for a satisfactory option. But aspiration levels adjust to the difficulty of finding a satisfactory option, and the search procedure determines which satisfactory option emerges first. So procedure plays a major role in satisficing. Also, the motivation for satisficing is practical. Hence Simon classifies the rule to satisfice as a procedural rule.

Many decision theorists have an interest in modifying standards of rationality to better suit the cognitive abilities of humans. Skyrms (1990 , 2–3) advocates using Simon's theory of bounded rationality and a dynamic approach to decision making to articulate the connection between decision rules and game theory. Rubinstein (1998 , 4) advocates using models of bounded rationality, applied especially to human inferential activity, to revise the accounts of equilibrium game theory advances. (See Bicchieri, chap. 10, this volume, on game theory.)

2. Criticisms

Because economics concentrates on instrumental rationality, a common complaint is that its conceptions of rationality are too narrow. Irrational goals escape criticism. For example, if an insane agent is bent on self-mutilation, economic rationality may seem not

to block his path. Furthermore, economics' approach to rationality may seem cold and calculating, and thus neglectful of the value of emotion and spontaneity.
The charge that economics examines only instrumental rationality is an exaggeration. The view that one should promote self-interest assumes that rationality requires that goal, for instance. Also, economists advance certain social goals as requirements of rationality. Case and Fair (1992 , 21–23) list efficiency, equity, growth, and stability. These goals are assumed in the branch of economics known as economic policy, which is allied with social philosophy. Still, it is true that economics downplays the evaluation of basic goals. It may easily correct the impression that it is averse to their evaluation, however, by acknowledging the need to supplement its principles of instrumental rationality with additional principles of rationality for assessing basic goals. Moreover, economics may acknowledge the value of emotion and spontaneity without abandoning its approach to instrumental rationality. Its principles of instrumental rationality may be formulated to accommodate all values, not just the materialistic ones common in textbook examples. Because economics may respond to the criticisms mentioned without significantly changing its conceptions of rationality, I do not examine those criticisms further. Also, I put aside criticisms of economic rationality prompted by its extension from single agents to groups of agents. Arrow's Possibility Theorem (1963 , chap. 5) presents a problem for this extension. It establishes that there is no satisfactory general procedure for aggregating the rational preference rankings of a group's members into a rational preference ranking for the group. Although some theorists suggest revising standards of rationality for individuals to facilitate their extension to groups, it may turn out that the standards appropriate for individuals just differ from the standards appropriate for groups. Perhaps an aggregation of individual preferences by a market or voting mechanism just does not produce results rational in any collective sense more rich than efficiency. Given this possibility, the problem of extending standards of individual rationality to groups does not yield straightforward criticisms of those standards.
The following three subsections review criticisms that target specifically the three conceptions of rationality section 1 presented.

Commitments

Many take exception to the view that rationality requires egoistic maximization of self-interest. They may concede that in trade, business, and markets rationality operates in a largely egoistic way. Still, they oppose applying the egoistic con
end p.387

ception of rationality to all walks of life. They resist the “commodification” of education, health care, citizenship, and family life. Pace rational choice theory in the social sciences, they think that egoistic rationality should be constrained to the economic realm and not extended to social and political institutions such as the family and the law.
Sen (1977) goes a step further. He argues against taking rationality to be the maximization of self-interest even in the economic realm. He observes that humans are

not motivated solely by self-interest, even in economic transactions where Edgeworth's assumption may seem to be a reasonable simplification. He observes that people are sometimes motivated by commitments to others—for example, to family, community, or class—and holds that such motivations are rational. Taking account of commitments, he claims, helps economists treat phenomena that fall squarely within their discipline's province, for example, the funding of public goods (such as roads, street lighting, and national defense) and the organization of a firm (including relations between labor, management, and shareholders). Duty may explain a citizen's honest revelation of a public good's benefit to him and may explain a worker's productivity in the absence of a comprehensive incentive system. Sen proposes revising economics' conception of rationality to allow for the commitments parents have to the welfare of their children, professionals have to the welfare of their clients, and citizens have to the welfare of their communities. Being motivated by commitment is compatible with a reasoned assessment of acts, and so is compatible with rational behavior in a broad sense (1977, 342–44).

Commitments to Teams and Plans

If utility is defined in terms of choices and the standard of utility maximization is just a matter of making consistent choices, then that standard of rationality is open to the charge of being too weak. As Sen (1987 , 70) points out, an agent may consistently choose to frustrate his goals. Such behavior is irrational despite being consistent. Similar criticism arises if utility is defined in terms of preferences and the standard of having preferences that follow utilities is just a matter of having consistent preferences. An agent comparing risky options may have consistent preferences that are as if he assigns probabilities 40 percent and 60 percent to an event and its complement, respectively, when in fact his probability assignments are the reverse. Such preferences are irrational despite being consistent.

In view of these powerful criticisms, this section examines utility maximization given a concept of utility that lets utility guide preferences and choices. It takes utility as a measure of strength of desire. Economists often have this concept of utility in mind when they take utility as satisfaction because they imagine that

satisfaction arises from attainment of any goal, including an altruistic one. Taking utility as degree of desire, utility maximization may still be regarded as a type of consistency between desires and choices, but it is a demanding type of consistency.

Utility maximization so conceived faces some objections. Some theorists object that it makes sense only when desires are quantitative and yield comparisons of options. When desires are not quantitative and options are incomparable, it does not apply and supplementary principles are needed. See Chang 1997 . Imprecision and incomparability certainly limit application of the principle of utility maximization but do not challenge its credentials where it applies. So I move on to another, more biting objection.

Some theorists object that utility maximization, even with utility broadly conceived, ignores certain types of commitment. According to one version of the criticism, it ignores commitments to teams of agents. According to another version, it ignores commitments to plans.

The first complaint is that maximization of personal utility, even if it accommodates altruistic goals, is too thin a conception of rationality to handle commitment to a team's goals. Consider, for example, a team of two altruists facing a coordination problem. Suppose they may coordinate in two ways, one of which is superior. Each altruist aims to participate in the coordination scheme the other follows. But what steers the pair to the superior coordination scheme if for some reason they cannot communicate? Each may identify with the team and pursue the team's goals and so achieve the superior form of coordination. According to Bacharach (forthcoming), such familiar behavior calls for standards of rationality going beyond maximization of personal utility.

The second complaint is that utility maximization does not provide for the instrumental value of commitments to plans placing restraints on utility maximization. A plan involves a commitment to a series of acts, perhaps including cultivation of a character trait or a behavioral disposition. Some steps in a plan may not be utility maximizing but may be justified by the benefits of executing the whole plan. Utility maximization may be applied to long-term patterns of behavior rather than to every act as the occasion for it arises. Acts may be justified by being part of a utility maximizing plan rather than by being utility maximizing themselves. Restraint in maximizing utility at a moment may serve the goal of maximizing utility over a lifetime.

Gauthier (1986 , chap. 6) advocates constrained maximization—roughly, acting on behavioral dispositions that are beneficial although they may call for nonmaximizing acts. Consider, for example, two people in a Prisoners' Dilemma, a decision problem in which cooperation by both benefits both but each has an incentive not to cooperate whatever the other agent does. Utility maximizing individuals fail to achieve cooperation. However, Gauthier points out that someone disposed to cooperate in a Prisoners' Dilemma profits from having the disposition if she is in a community of individuals who cooperate only with a person

having that disposition, even though her not cooperating still maximizes utility for her. According to his view, it is rational to cultivate the disposition to cooperate in those circumstances, and the cooperation to which the disposition leads is rational in virtue of the rationality of the disposition.

McClennen (1990 , chap. 1) considers sequences of acts. He entertains cases in which a sequence maximizes utility although acts within the sequence do not maximize utility. In such a case his view is that the sequence's utility maximization makes rational the nonmaximizing acts in the sequence. A rational agent resolutely carries out the sequence of acts despite the incentives for deviation. For instance, suppose that you can win five dollars now if you turn down a dollar offered later. You resolutely form the plan to turn down the dollar offered later and are rewarded five dollars. Later you act on your resolution to decline the dollar offered although you already have been paid five dollars and would not lose that money if you were to accept the dollar. According to McClennen,

acting on the resolution is rational despite not maximizing utility. (See McClennen, chap. 12, this volume, on rules.)
Bratman (1987 , chap. 2) draws attention to reasons that planning itself generates. Adopting a plan, he says, gives an agent a reason to follow the plan, a reason independent of utility maximization. If a driver is indifferent between two routes to his destination but settles on one, his having settled on that route generates a reason to take it. The reason does not depend on utility. Because of it, taking the other route is irrational although taking that route does not lower utility.

Bounded Agents

The principle to satisfice rather than optimize draws attention to human limits and the costs of making decisions. Although suggestive and intriguing, it has not replaced principles of optimization in economics. It needs further articulation. An agent's aspiration level should be carefully regulated so that what counts as a satisfactory option adjusts appropriately to circumstances. In some cases of impending disaster, the level should drop dramatically to prompt a fast decision. In well-studied, routine cases where optimizing is easy, it should rise to the maximum attainable. Even after regulation of an agent's aspiration level, the principle to satisfice is still incomplete. Since the search for a satisfactory option determines which satisfactory option is found first, and so which option the principle recommends, the search procedure should be fully specified. Also, cases in which an agent finds multiple satisfactory options simultaneously require a supplementary selection principle.
end p.390

Satisficing amounts to optimizing if an agent's aspiration level reaches an option at the top of her preference ranking, or if she is indifferent between satisfactory options. Some ways of articulating the principle to satisfice may make satisficing equivalent to a form of optimizing in all cases. Satisficing may be regarded as an optimization procedure whereby agents maximize utility but figure into utility calculations the cognitive costs of making decisions. Given limited cognitive powers, the costs of protracted deliberations are high, so shortcuts optimize all things considered. Raiffa (1968 , ix–x), for example, claims that his decision method is practical and takes account of human limits, although it aims at optimization. It registers the cost of deliberation when evaluating decision-making strategies and looks for an optimal strategy including that cost. Such optimization preempts the reasons for satisficing. Because it takes account of human limits, alternative principles of rationality are unnecessary. From the perspective of such theories, agents are bounded, not rationality, and satisficing, where defensible, agrees with rather than rivals optimizing. [2]
Simon recognizes that satisficing may be formally equivalent to optimizing if decision costs are included in evaluations of options. He regards forms of optimizing that take human limits into account as forms of satisficing or forms of procedural rationality (1982 , 435). According to him, emphasizing procedural rationality rather than substantive

rationality nonetheless makes a practical difference to problem-solving strategies (1982 , 417–18). This defense of satisficing takes both satisficing and optimizing as decision-making procedures rather than as standards for evaluating decisions that an outside observer may apply after the decisions have been made. This distinction also arises in another defense of satisficing.

Mongin (2000 , 95–104) argues against the attempt to reduce satisficing to optimizing. He claims that because of decision costs, optimizing is not always rational. Optimizing, he says, requires a decision to apply a principle of optimization, and optimizing in that second-order decision requires another, third-order decision to apply a principle of optimization, and so on. Thoroughgoing optimization therefore generates an infinite regress of decisions and therefore incurs prohibitively high decision costs. This objection arises, however, only if optimization is taken as a decision procedure rather than as a standard of evaluation. Meeting an optimizing standard of evaluation does not require a decision to apply an optimizing principle and so does not generate an infinite regress of decisions.

Whether satisficing rivals optimizing thus depends on the interpretation of the principle to maximize utility. The principle may be advanced as a procedure for decision makers to follow, or as a standard for evaluating decisions. Since economic rationality traditionally addresses standards of evaluation for decisions rather than decision-making procedures, I take utility maximization as a standard of evaluation and put aside the issue of rational decision-making procedures. Under this interpretation, utility maximization more easily addresses the concerns of proponents of satisficing. Taking the principle as a standard of evaluation is a way of recognizing that cognitively limited agents are unable to make every decision using optimizing procedures.

Slote (1989 , chap. 1) argues against optimizing and for satisficing on grounds independent of procedural advantages. He thinks that relentless optimization is immoderate. The virtue of moderation suggests satisficing, or being content with "good enough." Slote holds that it is rational, out of moderation, to reject the optimal in favor of the good enough—that is, to decline knowingly an optimal option (1989 , 12). In such cases optimization is practical but satisficing is proposed nonetheless.

Moderation is a virtue, but optimization may respond to its value. Consider an agent who appreciates the value of moderation. His utility assignment is affected by its value. If he must decide between two options and prefers one all things considered, including the value of moderation, then rationality requires its selection even if the other option is satisfactory. Moderation gives no reason for acting contrary to preference. Optimizing is not immoderate in this case.

In general, arguments for departures from utility maximization tend to maintain that some acts are preferable to utility-maximizing acts. The arguments may be rebutted by letting utility encompass the good served by those acts so that they count as utility maximizing. In the case of principles of bounded rationality such as satisficing, recognition of cognitive limits, decision costs, and other factors motivating the principles suggest broader accounts of utility that include those factors. Those accounts tailor utility maximization to the abilities of cognitively limited agents. Similarly, broader accounts of utility may tailor utility maximization to the values of moderate agents. I conclude that the arguments for satisficing do not produce a genuine rival to the principle of utility maximization if the latter adopts a broad interpretation of utility.

3. Circumspect Maximization

Some critics of economics' ideas about rationality promote alternative approaches to rationality. Gert (1998 , 39, 83–84) takes a deontic approach according to which an action is rational as long as it does not transgress commonsense principles such as those prohibiting gratuitous self-infliction of pain. Anderson (1993 , chap. 2) takes an action to be rational if it expresses rational values such as the value of health. Suppes (1984 , 187–203) takes an action to be rational if it issues from reasoning.
From a theoretical perspective, rather than make a fresh start, it is more
end p.392

appealing to revise economics' claims about rationality to meet criticism and so extract if possible the truth they contain. It is more fruitful theoretically to build on their foundation than to start from scratch. Let other approaches' insights supplement rather than supplant economics' approach to rationality.
I revise and defend utility maximization, with utilities for an agent taken broadly so that they register all her goals, including altruistic ones, and with utilities not defined in terms of preferences but rather introduced by utility theory in a way that makes them merely inferable from preferences and so still able to justify preferences. I take utility maximization as a standard of evaluation for decisions and advance it only for cases where options have utilities and so are comparable. The following subsections offer a brief defense of utility maximization. In *Decision Space* (2001) I more thoroughly examine it and methods of analyzing an option's utility.

Comprehensive Utility

My way to meet criticisms of utility maximization is to let utility swallow up the reasons advanced for deviations. An option's utility responds to any reason to adopt the option. Any reason to adopt it is a reason to prefer it to other options and so to assign it a utility higher than theirs. Resistance to this defense of utility maximization argues that a broad interpretation of utility makes utility maximization trivial or even unfalsifiable. It may seem that utilities can always be assigned so that a person's choice maximizes utility. However, to make the interpretation of utility broad is not to license an arbitrary assignment of utilities. They must still represent a person's preferences. Putting aside operational definitions of preference and adhering to the ordinary, everyday concept of preference, a preference is not definitionally linked to a choice. It is possible for an agent to act contrary to a preference. This possibility ensures the possibility of an agent's failure to maximize utility. A principle that advises utility maximization is therefore not vacuous.

I apply my defense of utility maximization to criticisms appealing to commitments. If a person becomes committed to the welfare of others, then he gives acts that promote their welfare a higher utility rating. Commitment to others is not contrary to utility maximization. Comprehensive utility covers the value of honoring such commitments. Does maximization of utility thwart some community values even if utility is comprehensive? Utility maximizers take many steps toward cooperation and its benefits. In current interactions they take account of the consequences of their choices on the behavior of others in future interactions. If an agent's cooperation with others now encourages them to cooperate with him later, he has an incentive
end p.393

to cooperate now. Also, societies, once established, may construct institutions to promote cooperation. Governmental authority may require cooperative measures that benefit all. Cooperation gathers momentum if a society's members have some measure of altruism that reduces incentives contrary to cooperation. Individuals with communitarian values may simultaneously maximize utility and promote social goals. Social dynamics, as Skyrms (1996 , chap. 3) explains, may account for the evolution of cooperation without novel standards of rationality.

Does utility maximization nonetheless fall short of optimal forms of coordination in some cases where team members cannot communicate? There are two ways of taking this question. One asks whether adjustments in utility assignments may yield the optimal form of coordination. The answer is yes. Identifying with a team may make an agent's utility assignment conform with the team's. Then the agent's utility maximizing behavior yields the same result as novel principles of rationality concerning team action.

Another way of taking the question asks whether, given the utility assignments agents typically have, utility maximization may yield coordination. The answer is again yes, given time for social dynamics to operate. Skyrms (forthcoming) shows that rational people seeking an optimal form of coordination may succeed by learning to associate with those seeking the same form of coordination. Attending to association building, people may optimally solve coordination problems without reliance on novel rules of rationality.

Does recognition of the value of planning require provision for deviations from moment-to-moment utility maximization? Consider the following case in which an agent's lifetime utility apparently suffers because he maximizes utility moment to moment. Suppose that an agent is surrounded by people who pester him. He never retaliates when attacked, however, since such retaliation brings no benefit after he has suffered an attack. Those committed to the policy to retaliate if attacked despite the costs do better over a lifetime than he because of the deterrent effect of their policy.

Although in this case nonmaximizing retaliation is rewarded, circumstances do not make such retaliation rational. If irrational retaliation is rewarded because people give berth to a person who retaliates irrationally, then a rational person cultivates that irrational behavior, as Pindyck and Rubinfeld (1989 , 479) observe. The principle to maximize utility acknowledges the rationality of cultivating that behavior without conceding the behavior's rationality. Rational agents cultivate dispositions and character traits that maximize lifetime utility. In environments where irrationality is rewarded, they may

cultivate dispositions and character traits that generate irrational behavior. This possibility does not demand revision of the principle to maximize utility. It remains the standard of rational behavior even in harsh environments.
Finally, reconsider commitment to plans. The reasons for such commitments provide reasons to honor them. Sticking to a plan has benefits despite the attrac
end p.394

tions of deviation. Failure to stick to a plan may thwart one's objectives. The constant search for improvements has cognitive costs. Agents with limited cognitive ability profit from resolutely following plans instead of calculating at each moment the act that maximizes utility. In cases where it is rational to follow one's plan, the cognitive costs of alternative strategies make following the plan utility maximizing. Recognition of cognitive limits brings utility maximization into alignment with rational planning procedures.

Recognition of Limits

Let me return to the objection that utility maximization is too lofty a standard for humans. My way of meeting this objection is to acknowledge that nonmaximizing behavior may be rational unless an agent and her decision problem meet certain idealizations. According to the idealizations, the agent is cognitively perfect and circumstances for making a decision are perfect. The agent has been perfectly rational up to the time of the decision and will be perfectly rational subsequently, so that her decision need not compensate for irrationality elsewhere.
Because economic theory uses rational behavior as an approximation to actual behavior, as Harsanyi (1977 , 16–19) and Sen (1987 , 68) observe, it may assume for simplification that rational humans maximize utility, although, strictly speaking, it advances utility maximization as a necessary condition of rationality only for ideal agents in ideal circumstances. Furthermore, despite the underlying idealizations, utility maximization may guide our decisions. It expresses a goal of rationality for them. Decisions that fall short of the goal are open to criticism even if nonideal conditions provide good excuses for falling short and so blunt a charge of irrationality. It may be a goal guiding decisions and decision preparations even though meeting the standard it expresses does not require adopting it as a decision procedure.
A standard method of generalizing a theory that relies on idealizations is to rescind an idealization and then adjust the theory's principles for its absence. The section on satisficing, for example, discussed a way of revising the principle of utility maximization to accommodate cognitive limits. It suggested making an option's utility take account of decision costs. A way of doing this is to take an option as a possible decision. That is, if I am deciding whether to go to the store, one option is the decision to go to the store, not just the act of going to the store. The decision's utility includes the decision's costs—for example, the time and effort invested in forming the intention to go to the store. These costs lower the decision's utility and influence its place in a utility ranking of options.

To illustrate another generalization of utility maximization for nonideal cases, consider a decision problem in which options lack quantitative utilities because
end p.395

desires are imprecise. As a result, the principle of utility maximization does not apply. Suppose, however, that despite not being quantitatively compared, the options are comparable. Then a generalization of utility maximization, formulated by I. J. Good (1952 , 114), applies. It involves what I call a *quantization* of beliefs and desires. This is just a probability and utility assignment that agrees with beliefs and desires. For example, if an agent wants a car more than a motorcycle, but the intensity of his preference is indeterminate, then one quantization may represent that preference as twice as intense as another does. Good's principle says that a rational decision maximizes utility under some quantization of beliefs and desires. According to it, utility maximization indirectly constrains rational choice even when beliefs and desires are imprecise.

Although removing idealizations and revising the principle of utility maximization to accommodate greater realism is a challenging project, the two adjustments sketched show that progress is possible. In *Realistic Decision Theory* (Weirich forthcoming) I pursue this project.

Generalization to Self-Support

Besides generalizing the principle of utility maximization to handle cognitive limits, it is important also to generalize it to handle certain technical problems. Generalization brings out more clearly fundamental principles behind traditional forms of utility maximization. Suppose an agent confronts an infinite number of options and for each option another has greater utility. Then it is impossible to maximize utility. The principle of utility maximization needs generalization for such cases. Also, utility depends on subjective probability and thus information, and adopting an option itself supplies information. What is the appropriate body of information to use when computing an option's utility? It seems that an option's utility should be computed with respect to the information that it is adopted. But since different options then have utilities computed with respect to information differing about the option adopted, comparisons of options lack a common basis. Even if the number of options is finite, it may turn out that no option maximizes utility on the assumption that it is adopted. In terminology Jeffrey (1983 , sec. 1.7) introduces, it may turn out that no option is ratifiable. For example, suppose you are competing with another player in a game, and you have just two options. If your opponent is gifted at outwitting you, it may be that whatever option you adopt, its adoption is evidence that your opponent has foreseen and countered it so that the other option would have been better. The principle of utility maximization needs generalization for such cases, too.

In *Equilibrium and Rationality* (1998, chap. 4) I propose a generalization that handles both problems. The first step defines an option's being self-supporting in a way that makes self-support subsume having maximum utility in standard cases but does not

require having maximum utility in general. The next step shows that in every decision problem, even the problem cases, some option is self-supporting. The final step argues that, under appropriate idealizations, rationality requires adopting a self-supporting option.

Conditional Rationality

Refining the principle of utility maximization to meet objections also calls for a distinction concerning the scope of evaluation for a decision. We sometimes say that an agent's decision was rational given his beliefs but was irrational all things considered because his beliefs were irrational. Similarly, we sometimes say that an agent's decision was rational given the options he entertained but was irrational all things considered because he failed to entertain better options. The scope of a decision's evaluation may exclude some relevant factors or include all relevant factors. We may adopt more or less comprehensive standards of evaluation. Noncomprehensive standards may assess the rationality of a decision taking for granted some factor that comprehensive standards call into question.

If idealizations do not remove mistakes an agent brings to a decision problem, utility maximization is a standard of conditional rationality only. It noncomprehensively assesses an agent's decision, taking for granted his utility assignment. The nonconditional standards that form the basis of a comprehensive evaluation must consider how a rational decision compensates for mistakes. In some cases departures from utility maximization may be appropriate. For example, a comprehensive evaluation may urge a decision that fails to maximize utility if the agent's utility assignment is irrational—say, if it springs from irrational anger. A maximizing decision may be rational taking for granted the agent's utility assignment but irrational from a comprehensive perspective that does not take for granted that condition of assessment. Working out the interplay of conditional and nonconditional standards of rationality is an exciting research project.

NOTES

I thank Larry Alexander, Peter Vallentyne, and Xinghe Wang for helpful thoughts about my topic.

1. This distinction between procedural and substantive rationality differs from the
end p.397

distinction made in ethics. For that distinction, see Hooker and Streumer, chap. 4, this volume.

2. Sen (1997 , 763, 768–69) distinguishes optimization from maximization and holds that maximization best assimilates satisficing. I do not distinguish optimization and maximization here.
end p.398

chapter 21 LEGAL THEORY AND THE RATIONAL ACTOR
Claire. Finkelstein

1. Introduction

The application of rational actor theory to the law has been dominated by the law and economics movement. Law and economics began as a methodology for approaching legal questions; it suggested using the tools provided by economic theory to solve problems in manifestly economic fields like antitrust and taxation. But it quickly moved beyond methodology to acquire both a normative and a descriptive thesis of its own. The normative thesis is that the legal system should serve the goal of maximizing social welfare. The descriptive thesis is that the common law has developed according to this implicit economic logic, even if judges and other legal actors have not consciously tried to fashion legal rules with economic theory in mind. Law and economics thus regards law as a vehicle for maximizing utility, and it maintains that much of law can be explained as serving that function already. [1]

In what follows, I shall leave law and economic's descriptive thesis about the nature of legal rules to one side and focus on its normative claim instead. On the normative side, law and economics shares a theory of value with act-utilitarianism in moral theory: Both posit some felicific notion like utility, well-being, happiness, or pleasure as the *summum bonum* for human beings. There are, however, several important differences between the two theories. First, legal economists have a

tendency to treat the notion of utility more generically than do ethical utilitarians. They include in it everything that is a source of value to an agent, whether or not it confers pleasure or happiness. Thus the legal economist might treat a person's altruistic feelings, aesthetic sentiments, and moral commitments as part of his utility, since these are all things a person values. Arguably, this makes the notion of utility vacuous, since it is not clear what could falsify the claim that a rational actor seeks to maximize his utility. (The economist, however, actually embraces this result, since he constructs an agent's utility function externally, namely from the choices he observes that agent making, rather than on the basis of the agent's internal psychological state.) For purposes of argument, then, I shall assume a somewhat narrower conception of utility, namely one that roughly corresponds to the idea of personal well-being. [2]

A second difference between law and economics and act-utilitarianism has to do with the object of evaluation. While utilitarianism treats individual acts as the relevant units from which maximizing must take place, law and economics maximizes from the standpoint of legal rules. Law and economics thus proposes a two-tier theory: Legal rules should maximize social welfare, while individual acts falling under those rules need not be

themselves maximizing. In other words, law and economics is a kind of rule utilitarianism, where the relevant rules are legal, rather than moral, in nature.
A third difference will provide the focus for our discussion. Law and economics is generally accompanied by a descriptive thesis about human nature, one that ethical utilitarians do not normally share. Legal economists tend to assume that the utilitarian theory of value to which they subscribe goes hand-in-hand with what we might call "rational actor psychology." [3] According to this view, human beings are rational maximizers who reason instrumentally toward the attainment of their ends. One can understand why the legal economist thinks the utilitarian theory of value presupposes a rational actor model of human agency: If society's good consists in maximizing social welfare, it seems natural to suppose that individual good consists in maximizing personal welfare. But a moment's reflection should make clear that the psychological theory is neither presupposed nor entailed by the theory of value, since we need not think of the subjects of a utilitarian regime as capable of reasoning about their own good. We could, for example, ask what the best life for a cow would be and seek to maximize its utility by providing it with grassy fields and plenty of water. But we need not think cows capable of utility maximization on their own behalf, still less of anything like instrumental reasoning. The only requirement that social utility theory does impose on the creatures to whom it applies is that they be capable of experiencing pleasure and pain, since without this we could not meaningfully speak of their having any utility or well-being to maximize.
Not only is rational actor psychology not inherently linked to the utilitarian theory of value, but the two are actually in some tension with one another. The

tension stems from the fact that when we maximize *social* welfare, we do not necessarily maximize the welfare of each individual member of society. In most cases, maximizing social welfare will result in some people faring worse than they otherwise would, even though other people fare better. This fact produces a quick route to the central claim I wish to make: If human beings are assumed to be rational maximizers, there is no reason to suppose they will be social maximizers as well. [4] And this suggests that a theory like law and economics that subscribes to a utilitarian theory of value at the same time that it adopts a rational actor psychology has some explaining to do. It must have a way of showing why the utilitarian theory of value it proposes should recommend itself to creatures with the particular constitution of rational maximizers. Indeed, I shall argue that rational actor psychology more naturally leads to a contractarian than to a utilitarian theory of legal rules.
We must now back up, however, and take the long route to my conclusion. For the legal economist thinks he has a ready answer to the above concern. Unlike act-utilitarianism, law and economics does not actually require individual agents to maximize social utility. It is only legal *rules* that must maximize social utility, while individuals living under those rules need think only of their own good. Good legal rules will bring the incentives of individual agents into line with the requirements of social welfare. Thus individuals living under correctly crafted legal rules will maximize society's utility when they attempt to maximize their own utility. Utilitarianism and rational actor psychology can

co-exist in law and economics, then, because it is possible to adopt social rules that equip rational, self-interested agents with incentives to maximize social welfare. In what follows, I shall nevertheless take issue with this way of reconciling rational actor psychology and the utilitarian theory of value. I shall suggest that the tension between the two persists as a result of two central features of legal rules. First, I argue that legal rules must be created and administered by individuals who are themselves rational maximizers, and thus they must see themselves as personally advantaged by the rules they institute. Second, I argue that if human beings are rational maximizers, we ought to be able to justify a legal principle like utility maximization by showing that rules adopted in accordance with that principle would be in their interest. Yet there is reason to think that rational agents would not select rules that maximize social welfare. Indeed, as I shall argue, there is reason to suppose they would select rules premised on some sort of contractarian norm of agreement instead.

2. Rational Actor Psychology and Legal Actors

The first problem with the legal economist's picture arises because rational actor psychology must be presumed to apply to what I shall call "legal actors," namely people who create and implement legal rules. The tension between rational actor psychology and the utilitarian theory of value will thus resurface, since there is no reason to suppose that legal actors such as judges, jurors and legislators will maximize their utility by writing and implementing legal rules that maximize social utility. There would be no difficulty here if some other, perfectly utilitarian entity were to determine social and political arrangements for us. For in this case, like the cows about whose good *we* might reason, the motivations of the rule-makers need not mirror the motivations of the creatures to whom the rules would apply. If legal economists maintain that rational human beings would adopt and faithfully apply rules that maximize social utility, they should also be able to tell us why doing so should be in the interests of a rational legal actor.

Some legal economists have noticed the difficulty, and they have sought to bring their account of the incentives of legal actors into line with their general account of individual motivation. Consider, in particular, the attempts legal economists have made to explain judicial behavior. Robert Cooter, for example, argues that public judges are primarily driven by the desire for prestige, and this leads them to act "in such a way that they would be chosen to decide cases by litigants and their lawyers if choice were allowed" (1983 , 107). They will thus maximize their personal utility if they faithfully apply the rules as they are written to the cases that come before them. And since those rules are ideally welfare-maximizing, judges will maximize social welfare when they maximize their own welfare.

Richard Posner gives a slightly different explanation. He argues that the judicial utility function is based on the pleasure judges experience in voting on cases, since it is by rendering decisions that they have a sense of their own power. (1995, part 1, sec. 3). He analogizes this pleasure to the satisfaction members of the ordinary public receive from voting in elections—the sense of one's own importance that comes from the fact that one is entitled to *decide*. Consistent with this, he argues, is the fact that judges do not

particularly care about writing opinions. They see their function as residing in the decisions they render, rather than in the reasons they might articulate for their decisions. But given the job and salary security most judges enjoy, this does not explain why judges render decisions that adhere to legal rules, instead of deciding cases based on personal or idiosyncratic preferences. Why, that is, do judges care about the rule of law, when it does not maximize their personal welfare to do so?
end p.402

Posner's answer is that adherence to the rules of judging is part of what makes the activity of judging pleasurable, much in the way that adherence to the rules of chess helps to make chess pleasurable (129). Judges adhere to the rule of law because they get more pleasure from following the rules, knowing it is this that defines their role, and hence their power, than they would if they decided cases for reasons of personal aggrandizement. He writes:

The reason is not that judges have different utility functions from those of other people; it is that the utility they derive from judging would be reduced by more than they would gain from giving way to [various] temptations.It is the same reason why many people do not cheat at games even when they are sure they could get away with cheating. The pleasure of judging is bound up with compliance with certain self-limiting rules that define the "game" of judging.It is by doing such things that you know you are playing the judge role, not some other role; and judges for the most part are people who want to be—judges. (131)

We could presumably tell a similar story about legislators. People who become legislators enjoy constructing rules that maximize social welfare. They are "public-spirited," in the sense that their own utility increases to the extent they feel they are serving the public good. Indeed, carrying the parallel further, we might say that legislators derive their sense of their own power from their ability to create social legislation that incorporates their view of what the public good requires. For this reason, there need be no tension between the demands of individual welfare and those of social welfare. Presumably we could explain the behavior of jurors along similar lines.

There are several problems, however, with this attempt to reconcile rule of law values with the hypothesis that legal actors are rational utility maximizers. I shall focus on Posner's account of judging, but similar criticisms could be made of the extension of that account to the other legal actors I have suggested. First, like the expansive approach Kaplow and Shavell take to the notion of utility, the rational actor psychology Posner's analysis assumes makes the claim that human beings are self-interested maximizers trivially true. If, for example, an agent has a strong moral compunction not to lie, cheat, or steal, this account will purport to explain the phenomenon by saying that he derives more utility from adhering to moral rules than from breaking them, since only then can he think of himself as playing a moral role from which he derives pleasure. Once again, this suggests that there is no motivational state that would falsify the hypothesis that legal actors are rational maximizers. Of course moral agents probably do derive pleasure from behaving morally. But we cannot use that pleasure as itself explanatory of the person's behavior if we wish to make a nontrivial assertion that the behavior is motivated by self-interest.

Second, Posner's account maintains that adhering to legal rules has instrumental value for a rational judge. But no account of this form is likely to succeed,
end p.403

for it fails to capture the experience of judges in relation to legal rules. Judges do not usually approach legal rules instrumentally. Instead, they regard themselves as bound by those rules, and do not feel free to ignore them if their personal utility would be enhanced by doing so. On Posner's story, in other words, we cannot explain the possibility of true internal commitment to the rule of law, since we cannot tell an instrumental story about such commitment. [5]

Third, if Posner were right, judges would systematically violate legal rules when, in a given case, they would gain more utility from breaking the rules than from abiding by them. The widespread adherence to legal rules among judges, even in cases in which individuals would have much to gain from violating the rule, belies this suggestion. It is possible, of course, that judges receive *such* positive utility from following legal rules that it is nearly impossible for them to receive more from violating those rules. But, as I have suggested, this hypothesis is either trivial or highly implausible: It is trivial if we have no independent measure of a judge's utility than what we discover he happens to choose, and it is implausible if the notion of utility is taken seriously as an independent variable, and the claim is that following legal rules confers such great personal pleasure that no bribe or other extra-legal inducement could outweigh it.

I thus return to my suggestion that insofar as legal rules and institutions must be created and implemented by rational agents, there is a tension between the utilitarian theory of value the legal economist espouses and the rational actor psychology to which he subscribes. For the theory to work, legal actors must be shown to be personally benefited by adherence to legal rules, and the few attempts that have been made to show this seem unpromising.

3. Rational Actors and the Social Contract

The second problem with the legal economist's proposed reconciliation of rational actor psychology and the utilitarian theory of value requires us to look back to an earlier point at which general principles for the establishment of legal institutions would be selected. What we must ask is what principle or set of principles putative members of a legal regime would settle on as the basis for constructing the legal institutions under which they must live. That is, what kind of legal institution would individual maximizers see as most in their interest to adopt? Broadly speaking, there are three possibilities: They could adopt institutions that follow the principle of utility maximization; they could establish institutions based
end p.404

on specific deontological commitments; or they could adopt legal institutions premised on the principle of mutual benefit. In other words, rational individuals might choose to make their legal institutions utilitarian, deontological, or contractarian. In what follows, I shall attempt to show why I think rational actors would not select the principle of utility as the theory of value for legal rules. I shall claim that they would also be unlikely to select deontological norms to structure legal rules. I conclude that rational agents selecting basic legal institutions are likely to select substantive norms of contractarian agreement. I shall attempt to demonstrate these claims by considering two examples of possible legal institutions.

First consider a somewhat simplified version of an accident law regime proposed by Kaplow and Shavell (2002, III. C). Suppose rational actors are given the choice between two different regimes: allow people to drive and assign liability for resulting accidents according to fault, or allow people to drive but let losses lie where they fall. (A third option would be to disallow driving entirely, but for present purposes I will assume this option would not be adopted.) In the first regime, a victim could sue an injurer and recover damages for his losses if the injurer was negligent in causing the victim's injuries. In the second regime, no victim could sue any injurer. Which regime would rational actors choose? Suppose that each person could expect to be both an injurer and a victim exactly once in his life, so that the risks imposed between injurers and victims were "reciprocal." In this case, Kaplow and Shavell argue, the parties would be indifferent between being able to sue for damages and leaving everyone to cover his own accident expenses. Now suppose there are high administrative costs involved in running a liability scheme, so that it is expensive for society to adjudicate and administer liability suits brought by accident victims. In this case, rational agents would choose to let losses lie where they fall, since the savings in administrative costs constitute a surplus that can be divided among the parties. The no-liability regime would leave at least one, and possibly all of the parties, better off than they would be under the liability regime—that is, the no-liability regime would be pareto superior to the liability regime. Thus, they conclude, the claim that a tort liability scheme is a required part of corrective justice should be rejected. Rational agents would not choose to equip themselves with a fairness-based remedy when a different legal scheme would leave at least one person better off and no one worse off.

Now consider the more realistic "non-reciprocal" case, the case in which the parties do not know when or whether they will be injurers or victims. In the absence of more specific knowledge of their situation, the parties should be indifferent between this case and the case in which the risk-taking is reciprocal. If there are high administrative costs to running a liability regime, Kaplow and Shavell once again argue, rational agents would choose the regime that leaves losses where they fall, since that regime would again be pareto superior to the

more expensive liability regime. As a general matter, Kaplow and Shavell assert that rational agents who do not know their ultimate positions under the regime they select will maximize their expected utility by selecting the regime with the highest social return. They would never choose a legal regime on grounds of "fairness." Any such regime

could leave *everyone* worse off, since even the worst off group in an unfair regime might have a greater absolute share than they would in a regime where returns are distributed fairly.

For this reason, Kaplow and Shavell may be correct that rational agents would not choose to equip themselves with a legal regime predicated on norms of fairness, and so to dispense with the suggestion that rational agents would select deontological norms *as such* to govern specific legal institutions. It hardly follows, however, that they would adopt the accident regime dictated by utilitarian norms. For the fact that a rational agent has a higher expected utility from a given legal regime does not *necessarily* mean that he would choose that regime, even if he is choosing strictly on the basis of self-interest. To see this, let us make use of an assumption that is common among legal economists, namely that a person who engages in risky behavior in some way benefits himself at the expense of others. The no-liability regime is thus one in which injurers improve their own position without having to pay for it, and victims are harmed without receiving compensation. Individuals living under such a regime are enrolled in a lottery, in which the injurers are the winners and the victims are the losers. By contrast, the liability regime would be one in which the parties' post-accident welfare would be restored to its pre-accident state through court-ordered damage awards. In such a regime, there would be no winners and losers once compensation was paid. Now the question we might ask is: Would rational agents prefer the regime in which they gamble, hoping they will be non-liable injurers rather than uncompensated victims? Or would they prefer the regime in which they are assured of their pre-accident, baseline welfare, but in which they do not particularly stand to gain if they are not injured?

In answering this question, let us assume that society would not allow people to drive without adopting one or the other of these liability regimes. That is, the "pre-agreement baseline" is the third option I mentioned above, namely that people do not drive at all. Arguably rational agents would prefer not to take a gamble that would potentially give them a lower payoff than they would receive from their *ex ante* baseline welfare. Indeed, their preference for protecting their baseline might be sufficiently strong that it would overcome even a higher expected utility from the regime involving a gamble. Rational agents, in other words, might be particularly disinclined to gamble with their baseline welfare, and be willing to forego the potential for even greater gains to protect that baseline. Why might this be the case? Consider how an agent contemplating an agreement for a future accident regime might reason. He might evaluate the rationality of the agreement by asking himself whether he would later be glad he entered that agreement or whether he would regret having done so. It is plausible to suppose that he would think himself well-served by the agreement if he is better off under the terms of the agreement than he would be had he never entered into the agreement in the first place.[6] That is, even if an agreement would require him to perform individual actions that leave him worse off, the agreement as a whole should leave him better off than he would be had he never entered into it. Let us call this condition the "benefit principle."

The requirement that an agent must receive a net benefit is a plausible threshold criterion for rational agents to apply in assessing options regarding the basic institutions of their society. The idea is that rational agents would not endorse basic institutions that they expect would leave them worse off than they would be in the absence of that institution. If rational agents did indeed apply such a test, they would not simply be concerned to

maximize their utility in selecting basic social institutions. Their interest in maximizing utility in this context would be subject to a constraint, namely that no matter how much expected utility an institution conferred ex ante, each agent should be reasonably confident that he would fare at least as well under that institution as he would in its absence. [7] The benefit principle should not be treated as a general test for the rationality of all agreements, plans or courses of action rational agents might adopt. For as a general condition of rationality, the principle would be much too strong: it would appear to rule out the rationality of homeowner's insurance, gambling (no matter how favorable the odds), and stock market investment. [8] But I am suggesting that such a strong condition would be rational with regard to agreements that govern the basic structure of society. Since rational contractors do not currently know the life choices they will be making, nor the kinds of preferences they are likely to have, they cannot count on ordinary calculations of expected utility to adequately protect their interests.

Would the agreement to establish the no-liability regime satisfy the benefit principle in the non-reciprocal case? Under the no-liability regime, if a person happened not to become injured in an accident, he would regard himself as better off under this regime than he would be had he never agreed to that regime, since he would be able to enrich himself by engaging in risky behavior without paying the "price" for it. But he might turn out to be a victim without ever being an injurer. In such a case the no-liability regime would leave him worse off than he would be if he had never agreed to this regime, since his welfare would then be substantially below his pre-agreement baseline, and he would receive no compensation. Since the rational agent does not know which of these scenarios will occur, he cannot be sure that the agreement to establish the no-liability regime is one he will be glad to have entered. The no-liability regime thus appears to violate the benefit principle.

What about the liability regime? No matter what position he happens to find himself in, an agent will regard himself as better off under the agreement to

establish the liability regime than if he had never entered into that agreement in the first place. For if he is an injurer, he will have to pay for the injuries he inflicts, but he will regard himself as better off for the opportunity to drive negligently, paying the price, than if he were not able to do so. (He presumably values the risky activity more than the discounted amount he must pay in compensation, otherwise he would not engage in the activity.) And if he is a victim, he is restored to his pre-accident baseline welfare, at the same time that he has the opportunity to engage in risky activities as a potential injurer. There is thus some reason to think the rational agent would prefer the liability regime. Indeed, he might prefer it sufficiently that he would be willing to accept it even if his *expected* utility were higher under the no-liability regime, as Kaplow and Shavell suggest. And if this is so, then we cannot assume that rational agents would systematically select the utility-maximizing solution.

In saying that rational agents would select a legal regime that satisfies the benefit principle, I have not uniquely identified the legal regimes that rational agents would adopt. In theory, there might be many possible legal regimes that satisfy that principle. My claim is only that rational contracting agents would reject any legal regime that failed

the benefit test, and so we have reason to think they would not consistently choose the legal regime that maximized social welfare. In a fuller account, one would need to specify some further principle of selection that would allow the parties to choose from among the various eligible regimes. Presumably that principle would be consistent with contractarian assumptions. [9] My suggestion, then, is simply that rational agents would select the benefit principle as a minimum necessary condition for judging the acceptability of any agreement regarding legal regimes into which they might enter.

Let us now apply the framework we developed in considering the accident law regime to a second question of legal policy, namely the death penalty. Assume that the death penalty has deterrent value over and above life in prison without parole. Would rational agents choose to include it in their punishment regime? If potential criminals are rational actors, the same level of deterrence can be accomplished either by combining a high probability of detection with low penalties, or by combining a lower probability of detection with more severe penalties. The legal economist therefore recommends that penalties be made as severe as possible, in order to minimize the use of costly resources in detecting and preventing crime. In general, the severity of punishment can be increased at less added cost than can the chances of detection. Thus economic analysis seems to point in the direction of using death as a penalty, if, as its proponents claim, the death penalty has greater deterrent value than life in prison without parole.

But would rational agents setting up an institution of punishment select the form of that scheme by applying the principle of utility maximization? There are two reasons to think they might not. First, notice that the legal economist's argument places no upper boundary on the level of acceptable punishment. Sup

end p.408

pose, for example, that torturing a person prior to executing him would increase the deterrent efficacy of the death penalty significantly. By the legal economist's lights, we should then opt for torture, since it would enable us to lower the amount spent on detection. It is probable, however, that the parties to an original social contract would treat certain kinds of very harsh penalties as off-limits, no matter what the utilitarian gains. The legal economist, by contrast, famously has difficulty explaining why certain morally impermissible punishments are unacceptable. Second, there is no reason to suppose that the person executed must be guilty of a crime in order for the death penalty to have deterrent efficacy. The legal economist's argument seems to sanction executing innocent agents. It is true that others will not be deterred from committing crimes unless they see at least a significant number of those who have committed those crimes receive a stiff enough sentence that they would rather abandon their plans than risk the associated punishment. But there is no need for the person punished actually to have committed the crime, strictly speaking. All that is really necessary is for the public to believe that he has done so.

Now it might seem that the only way parties to a social contract would be able to eliminate such implications is if they were able to agree directly on a set of deontological principles that would rule out torture and sacrifice of the innocent. But it is unlikely that putative members of society could reach agreement on substantive moral principles in this way. Thus once again I suspect that Kaplow and Shavell are correct to think that

rational agents would not choose to govern their institutions according to specific deontological norms. The benefit of seeking to establish the basic structure by agreement among rational agents is to accommodate the fact that individuals have different views of the good and so cannot reach substantive agreement. How, then, would the parties to a social contract determine their principles of punishment?

To determine whether rational agents would opt for the death penalty, let us once more apply the benefit principle, and ask whether a rational agent would regard himself as better off once he is subject to the death penalty, having lived with the benefits of a regime in which the higher deterrent effects of the death penalty are available, than he would have been in its absence. Does he prefer his life with the death penalty, now that he is subject to it, or the life he would have lived without that particular penalty? Presumably he would almost always prefer his life without the death penalty than with it, since he will not regard himself as better off once he is executed than he would have been had he never lived in a regime with the death penalty. That is, the agreement to institute the death penalty would have left him worse off, on balance, than he would have been in the absence of the agreement. The death penalty thus seems to violate the benefit principle and rational agents would therefore reject it.

In this regard, the death penalty is like a kidney lottery society, in which each person has the option of entering into the following agreement: Should he suffer

dual kidney failure and require a kidney to survive, a member of the society will be chosen at random to supply a kidney in order that the victim of kidney failure might live. In such a case, the kidney lottery clearly passes the benefit test, since the person receiving the kidney is better off than he would have been in the absence of the lottery. But suppose he never does suffer kidney failure. Instead, someone else suffers kidney failure and our agent is chosen by lot to supply *him* with the needed kidney. The kidney lottery agreement would clearly fail the benefit test in that case, since the person chosen to supply the kidney will then wish he had never entered the kidney lottery. In this case, the lottery has turned out to be all cost and no gain. It is true that he might still benefit from the agreement, because should his one remaining kidney now fail, he will be the beneficiary of a lottery in which someone else must supply a kidney. But this is not a case of *net* benefit, since in the end he will be left with one kidney, which is what he would have had if he had never entered into the agreement in the first place.

Arguably we should think of the death penalty as a kidney lottery. An agent who must be put to death by the state is like the person who suffers kidney failure under the lottery agreement: while he has undoubtedly benefited from the agreement to have the death penalty thus far, the agreement has not conferred a net benefit on him, since the fact that he will be put to death will now more than counter-balance any benefits he has received from the agreement. But how do we know this? Perhaps in the absence of the death penalty, everyone would die a violent death at someone else's hands at a very young age. So being able to live longer, and then be put to death, does not *necessarily* leave one worse off than one would have been in the absence of the death penalty. But in all likelihood, the availability of other severe punishments, such as life in prison without parole, would provide most of the deterrent benefits rational agents would require. The

marginal deterrent benefits death provides over other available punishments would almost certainly be insufficient to outweigh the loss of one's own life.
One might object that the death penalty is distinctly unlike the kidney lottery, in that the decision to commit a crime is subject to choice, whereas suffering kidney failure is not. But once we allow that the state makes occasional mistakes in the administration of the death penalty, and that indeed it is quite impossible to run a death penalty without occasional loss of innocent life, the death penalty becomes precisely like the kidney lottery, since for an innocent person put to death, the loss of life is as much under his control as kidney failure. Moreover, even if we could ensure that no innocent people were put to death, there is reason to see even one's later decision to commit a crime as beyond one's control. This is because, from the *ex ante* perspective, a rational agent knows nothing of his future motiviations. The need or desire to commit a crime may be so compelling that his future self has no choice but to comply. Under either scenario, a rational agent is no more likely to enter a death penalty lottery than he would be to enter the aforementioned kidney lottery.
end p.410

These examples are intended to suggest that utility maximization is not the only or even the most appropriate theory of value for a legal system designed by and for rational agents. Rational agents have reason to prefer legal institutions organized around the principle of mutual benefit, since each agent can be sure he will benefit in such a scheme. Thus if we take rational actor psychology seriously as a starting-point for legal theory, there is reason to question the claim that legal institutions would be organized around economic principles. It is at least as likely that they would be organized around contractarian principles instead.

4. Does Wealth Maximization Solve the Problem?

In an article in the *Journal of Legal Studies* (1979), Richard Posner tries to distance law and economics from utilitarianism for precisely the reason discussed in the preceding section, namely that utilitarianism endorses the sacrifice of individual welfare for the sake of the greater good. He does this by suggesting that law and economics need not be premised on utility maximization after all. Instead, the legal economist suggests that legal institutions should seek to maximize *wealth*. Posner argues that wealth maximization better captures our intuitions about appropriate legal institutions. For example, a poor man who steals an expensive necklace he could not afford to buy may increase the total amount of utility in the world, since his wife, to whom he gives the necklace, may enjoy it more than the wealthy person from whom it was stolen. But the thief nevertheless decreases society's wealth, assuming he could not afford to purchase the necklace, since in an economic sense that means he does not value it. Thus arguably we can make sense of the law of theft in a wealth maximization system, but not in a system that seeks to maximize utility. It is worth revisiting the criticism I made of law and economics in

section 3, then, to consider whether rational agents would see their interests as adequately protected by a legal system that sought to maximize wealth rather than utility.

There is some initial reason to think that wealth maximization would be an improvement in this regard. Posner argues that unlike utility maximization, wealth maximization requires voluntary exchange, since wealth cannot exist without markets. And voluntary exchange requires consent. As a result, law and economics will arguably no longer sanction the sacrifice of a smaller number of individuals for the sake of improving the well-being of others. For if society wishes to make use of the body or the possessions of one of its members, it will have to purchase the right to do so from him. As Posner puts the point:

The great difference between utilitarian and economic moralityis that the utilitarian, despite his professed concern with *social* welfare, must logically ascribe value to all sorts of asocial behavior, such as envy and sadism, because these are common sources of personal satisfaction and hence of utility. In contrast, lawfully obtained wealth is created only by doing things for other people—offering them advantageous trades. The individual may be completely selfish but he cannot, in a well-regulated market economy, promote his self-interest without benefiting others as well as himself. (1979, 132)

The thief, in other words, will have to purchase the necklace from its rightful owner rather than simply taking it, since a nonconsensual transfer would bypass a voluntary market.

If Posner is correct, arguably a system predicated on wealth maximization would not permit institutions that left individuals worse off than they would be under their baseline welfare, since, as I have claimed, no rational agent would consent to such an institution. Thus it might be possible to acquire the benefits of a contractarian system within the framework of an approach that seeks to maximize social welfare, simply by measuring social welfare in terms of wealth instead of utility.

One difficulty with this suggestion, however, is that economic analysis conceived in terms of wealth maximization is incomplete. The reason is that the rights respected in a market system are allocated prior to the operation of the market. That is, economic theory cannot itself specify the initial distribution of rights, since the theory actually presupposes such a distribution. In the case of the necklace, for example, we were assuming that the old owner of the necklace has rights to it that protected his entitlement when we said that the thief must purchase the necklace rather than just take it. But that system of rights has not itself been specified in economic terms; we have no apparent justification in terms of wealth maximization for allowing the old owner to reject attempts to remove the necklace from his possession without his consent. As Ronald Dworkin has convincingly argued, if the thief *were* willing to pay more for the necklace than the owner would accept to sell it, wealth would be maximized if a dictator were simply to give the thief the necklace, without requiring him to pay for it (1980 , 196–97).

Posner argues against this point by suggesting that economic analysis is not indifferent to initial entitlements. There are two reasons for this. First, Posner argues that in general it makes sense to give each person the right to his or her own labor, since a person is likely to be more productive if he is his own master. In general, he thinks, there are economic reasons that bodies and labor should

be assigned to their “natural” owners, namely that natural owners are likely to value them most. Second, if there are transactions costs, an initial assignment of right to someone who values a good less than another person may be inefficient, since the parties may not be able to bargain for the exchange of the good. Thus Posner writes, “If transaction costs are positive, the wealth-maximization principle requires the initial vesting of rights in those who are likely to value them the most. This is the economic reason for giving a worker the right to sell his labor and a woman the right to determine her sexual partners. If assigned randomly to strangers these rights would generally (not invariably) be repurchased by the worker and woman respectively” (1979 , 125).

The argument, however, is not a good one. Why assume that natural owners are likely to value their own labor or bodies more than others? Posner gives no defense of this claim, and on the face of it, it sounds more like an intuitive or moral idea than an economic one. The factory owner may indeed value the worker's labor more highly than the worker himself, in strictly economic terms, because he has the initial capital to make a large amount of money from the worker's labor, much more than the worker could ever be paid. Even a law firm will bill an associate's time at twice what they will actually pay the associate. The same might be said of a woman and her body: a pimp might make more from selling her body than she could ever make from selling her own. By substituting wealth maximization for utility maximization, Posner thus seems to save law and economics from the morally unacceptable results that come from its traditional association with act-utilitarianism. But it is not at all clear that he is able to do this without assuming a system of rights to property and to natural bodily entitlements. And if that is so, then law and economics can provide only a supplementary theory of how goods allocated according to non-economic principles can be fairly exchanged once initial entitlements are already in place.

There is a second problem with the idea of wealth maximization, which Dworkin has also helpfully pointed out (1980 , 194–95). In order to be valuable, wealth must either be something of intrinsic worth—that is, an item worth having for its own sake—or it must be instrumentally valuable relative to something else that has value. Utilitarians think of utility (or happiness) as something worth having for its own sake. It is not valuable because it produces other things of value. Rather, all items that have value have it because they contribute to the total amount of utility. In this sense, the value of utility must be understood as self-justifying, since utility is the value in terms of which all other items of value are justified. If Posner is proposing that law and economics seek to maximize wealth instead of utility, he might be thinking of wealth in this same way: It is the ultimate value in terms of which all other items are valued. But this account cannot be defended, and Posner himself has admitted as much. In response to Dworkin's critique, he writes that wealth is a value because “it is conducive

to happiness, freedom, self-expression, and other uncontroversial goods” (1980 , 244). As Posner recognizes, if wealth is a value, it *must* be an instrumental value—something that is valuable because it allows one to acquire other items of value.

But if wealth is only instrumentally valuable, it is not at all clear why legal economists would want to maximize it. Is there any reason to think it would be better to maximize wealth than to try to maximize whatever *is* of intrinsic value directly? For example, assume in a given instance that we have reason to increase social wealth because we believe it will increase total societal happiness and not decrease any of the other goods we value. Is there any reason in such a case to seek to maximize wealth *rather* than happiness? Presumably not. And if not, why would we imagine that the benefits of wealth maximization change when maximizing wealth would maximize several other values as well? Moreover, there is a disadvantage to maximizing wealth as a proxy for maximizing other values: we do not know what quantity of these other values we will achieve if we just aim at wealth, whereas we can strike the balance we prefer if we aim at each value separately. Posner also suggests that by aiming at wealth rather than utility we can "thereby avoid[] certain well-known problems to which utilitarianism gives rise" (1980 , 245), presumably the moral problems we discussed earlier. But if we cannot aim at maximizing utility because it produces morally unacceptable results, then surely it does not improve matters to aim at wealth. For if maximizing wealth really does maximize utility, it will produce those same results, and if it fails to produce them, it can only be because we are not, in fact, maximizing utility.

Not only does turning to wealth instead of utility fail to solve the second of the two problems with law and economics we considered above, [10] but it makes particularly clear just how serious the first difficulty is, namely the problem concerning the motivations of legal actors. [11] Is there any reason to suppose that an individual legal actor who is primarily concerned to maximize his own wealth would prefer rules that maximize social wealth? Not especially. It is possible that a rational legislator or judge would believe that his own personal wealth would be maximized if society were maximally wealthy, but the connection between the personal wealth of legal actors and society's wealth is not a particularly reliable one. A legal system that allowed legal actors to take bribes and kickbacks, for example, would probably maximize the personal wealth of legal actors, but it would be unlikely to maximize social wealth. Thus the gap between rational actor psychology and social welfare seems particularly significant if we substitute wealth for utility in the maximizing conception of value.

end p.414

5. Conclusion

I have argued that if human beings are rational, self-interested agents, they are most likely to favor institutions premised on the principle of mutual benefit, rather than on the maximization of social welfare. In this sense, rational actor psychology seems at least as likely to lead to contractarian as to utilitarian legal institutions. What it means for a legal institution to be based on contractarian principles will differ depending on the nature of the institution. But unlike a utilitarian approach, I have argued, the contractarian approach to legal institutions precludes legal rules that leave some worse off than they would be without those institutions. It also precludes the creation of institutions premised on the substantive moral commitments of any particular group that desires to force

compliance with their own views on others whose views diverge from theirs. What characterizes legal institutions premised on rational agreement is that the individuals that are subject to them can feel that those institutions have been established according to principles they accept as advantageous, even when the specific workings of those institutions are not.

NOTES

My thanks to Michael Abramowicz, Michael Green, Peter Huang, and Leo Katz for comments on earlier drafts and to Geoff Sayre-McCord for helpful conversations.

1. Although the term “welfare” is often taken to be broader than the notion of utility, I shall use the two terms interchangeably.

2. I do not think that this assumption in any way affects my disagreement with the legal economist. My criticisms would remain even on the broader conception of utility he proposes.

3. This psychological claim is not to be confused with what I referred to above as the legal economist's “descriptive” claim about the nature of legal rules. While it is a descriptive claim, it is a generic one about human motivation, one that is not particular to the legal economist.

4. Individual maximizers will produce a pareto optimal state of affairs. But that is not necessarily a state in which social welfare is maximized, assuming that each person's utility counts equally.

5. For a discussion of this point, in connection with Hart's account of the “internal point of view,” see Siegel 1999 , 1581 .

6. I have here adapted a criterion of David Gauthier's (1990 , 1994) for determining when entering into an agreement would be rational for an agent. But I make quite a different use of this criterion than Gauthier does, since I am applying it to test the rationality of enforceable agreements. There is arguably no need for such a test in this

end p.415

context, because the agent can execute a plan he finds attractive on expected benefit grounds, without having to rely on his own willingness to perform a suboptimal action by way of compliance. Gauthier's test is a way for an agent to rationalize compliance without an agreement or plan that is not strictly speaking in his interest at the moment he must do his part. But arguably the criterion is a useful one for enforceable agreements as well. That is, we might think that the rational enforceable agreement is just the agreement the parties would have entered into in the absence of an enforcement mechanism. Enforcement, on this view, would be no more than a way of bringing somewhat irrational agents into line with the solution they would adopt if they were fully rational. I cannot, however, defend this account in any greater detail on this occasion.

7. Someone might argue that an individual utility function could be constructed whose maximization would satisfy the benefit principle, and that I have not therefore challenged the rationality of straightforward utility maximization at all. But the function being maximized would probably lack at least one of the features that economists typically impose on rational agency, namely transitivity, completeness, and continuity.

8. There are, however, ways to make the benefit more plausible as a general principle of rationality. As I have suggested elsewhere (Finkelstein 2003), it is plausible to think of

an ex ante chance of benefit as itself a benefit. And if this is so, then insurance and some gambles would be worthwhile: it would be rational to sign up for insurance, for example, if the ex ante chance of benefit were itself greater than the cost of the premium. Thus if the damage against which one is insuring is sufficiently great, and the chance that the damage will occur sufficiently high, then paying an insurance premium for an insurance policy that one does not end up needing nevertheless conveys a benefit.
9. One possible candidate would be David Gauthier's "minimax relative concession." See Gauthier 1986 , Ch. V. Another would be Harsanyi's approach. See Harsanyi 1977 .
10. See supra section 3.
11. See supra section 2.

chapter 22 RATIONALITY AND EVOLUTION

Peter. Danielson

The most striking fact about the relationship between evolutionary game theory and economic game theory is that, at the most basic level, a theory built of hyper-rational actors and a theory built of possibly non-rational actors are in fundamental agreement. This fact has been widely noticed, and its importance can hardly be overestimated. Skyrms 2000 , 273
This chapter will focus on this "striking fact," more generally: how game theory, the abstract theory of strategic interaction, supports new and powerful insights into the relation between rationality and evolution. First and most broadly, both rationality and evolution are optimizing processes working in a multiagent environment. I show in section 1 that the same modern methods that undermine naive evolutionary progressivism support the claim that rationality and evolution are isomorphic optimizing processes. This claim is striking, because, on their face, evolution and rationality are so different. While rationality tends toward a normative theory that fully applies only to extremely sophisticated, well-informed superagents, evolution manages to account for adaptation of the simplest organisms or even simple computer programs. The theoretical unification of these diverse phenomena is a major accomplishment for the relatively new field of game

theory. The second remarkable result of recent theorizing about these two processes is that they allow evolutionary and rational processes to be precisely compared. In section 2 I contrast a simple rational and evolutionary model and defend my focus on game theory, sketching the main concepts and variants of evolutionary game theory. The third interesting result is how the residual differences between evolution and rationality illuminate some crucial problems in the theory of rational choice: equilibrium selection and modular rationality. Setting various kinds of agents to play the same game under different information and selection regimes allows us to study precise differences.
The research surveyed in this chapter comes from a broad range of disciplines, including mathematics, economics, biology, philosophy, political science, psychology, and computer science and engineering. This is remarkable in two respects. First, the

fruitfulness of the theories that unify evolution and rationality is indicated by their appeal to researchers in so many disciplines. Second, this unifying approach to social science has spread quite rapidly. For example, in 1967 the *Encyclopedia of Philosophy* (Edwards 1967) drew no connection between evolution and rationality, but by 1984, Robert Axelrod captures the convergence of game theory and biology by remarking that the Prisoners' Dilemma was the *E. coli* of social science. Later, "evolution" doesn't appear in the index of Elster 1986 's survey of leading contributions to rational choice theory, while half of Katz's recent collection on *Evolutionary Origins of Morality* (2000) concerns game theoretic arguments. While this chapter will not be historical, I will try to catch the interdisciplinary flavor of the material by quoting some of the original contributors of insights into the connection between rationality and evolutionary processes. In the fourth section, I consider the interdisciplinary aspect of the research that produced these results, with special attention to the role of models and simulations as a powerful research language; I sketch some two-level models that combine evolution and rationality. Fifth and finally, I return to the issue of normativity and consider whether the game theoretic approach to evolution is likely to encourage the evolution of rationality or something else.

1. Two Optimizing Processes

Evolution and rationality differ in so many respects that it is remarkable that they could be thought of as nearly identical. As Jorgen Weibull (1998 , 3) points out, "A qualitative differencebetween evolutionary and rationalistic approaches is that while the second focuses on individuals and what goes on in their minds, the evolutionary approach usually instead analyses the population distribution of
end p.418

behaviors (decision rules, strategies). One could say that the selection process replaces the mental process of choice made by rational players in classical non-cooperative game theory, while the mutation process replaces the mental process of exploring the strategy set and strategies['] payoff consequences." Elliot Sober puts the general point forcefully: "Why bother to write about differences between two processes [deliberation and evolution] that are so obviously different?Deliberation involves a change in the composition of an *individual;* evolution effects a change in the composition of a *population*.Yet, in spite of these manifest differences, there seems to be an important isomorphism between the two processes[both are] *optimizing processes*." This isomorphism relates evolution to rationality via what Sober calls "*the heuristic of personification*: If natural selection controls which of traits T, A_1 , A_2 ,, A_n , evolves in a given population, then T will evolve, rather than the alternatives listed, if and only if a rational agent who wanted to maximize fitness would choose T over A_1 , A_2 ,, A_n " (1998, 408–9).
Incidentally, this formulation of the analogy illuminates the (so-called) naturalistic fallacy, which Moore (1903) deployed to criticize some of Spencer's arguments (1884)

from evolution to various values and which many have taken to block any normative significance for evolution (see Flew 1967). The generic fallacy is committed by arguments that attempt to output value content without a value input. However, we can see from the analogy above that neither rational nor evolutionary optimization need commit this error. Rational choice gets value out only by optimizing the value input captured by the agent's preferences. Conversely, what evolution optimizes is a value only if reproductive fitness is valued.

The analogy becomes richer when we move to the perspective of multiple agents, organisms, or species. Darwin (1859 , 489) concludes *The Origin of Species* with this optimistic argument: “And as natural selection works solely by and for the good of each being, all corporeal and mental endowments will tend to progress towards perfection.” According to Sober's interpretation, Darwin is arguing that since selection maximizes individual fitness, it thereby maximizes group fitness. Sober points to the background assumption for this argument and indicates what goes wrong:

The economists of Darwin's time tended to think that since a society is “nothing but” a collection of individuals, the society will maximize its well-being if each individual endeavors to maximize his welfare.Before one understands the structure of the tragedy of the commons, it perhaps sounds like a contradiction to be told that each individual, acting rationally and correctly foreseeing the consequences of his or her actions, will make a decision that leaves everyone worse off than he would be if some “irrational” choice had been made. [1] Perhaps it takes a rigorous decision theory on the one hand and the abstract picture of natural selection afforded by Fisher's theory on the other to bring the issues into focus. (Sober 1984 , 193)

end p.419

“Selection may be relied upon to be a `progressive' force of evolutiononly when fitnesses are frequency-*in*dependent” (185).

Hardin's tragedy of the commons model (1968) was designed to rebut straightforward optimization problem solving in strategic contexts. In Hardin's simple (but controversial; see Monbiot 1994) commons model, each of us rationally adds an additional grazing animal to a limited common field. However, the success of our strategies is frequency dependent. More grazing animals degrade the resource leading to poorer grazing for all. This is a many-player case of the problem seen most simply in the two-player Prisoners' Dilemma game. Each player's payoff depends on the number of players choosing that strategy; defecting alone pays the best, but when the other player defects as well, one's payoff drops to second worst. A classic real world tragedy of the commons is Gordon 1954 's model of the Canadian Atlantic cod fishery, which unfortunately has proven so accurate as a prediction that Atlantic cod is now practically extinct. Generalizing, the prevalence of environments with this interactive structure—now called social dilemmas—blocks the straightforward evolutionary route to perfection, or so it seems. In Dawkins's apt terminology, in these cases it is difficult to discern “God's Utility Function” (Dawkins 1996 , chap. 4).

However there are some endowments so well suited for all environments, even these social dilemmas, that their evolutionary development should not be undercut by this argument. While grazers' and fishers' *welfare* will not be maximized by evolution in a

commons, their *rationality* should be. Rationality, after all, is the perfection of just those abilities useful for exploiting any situation, including social dilemmas. To use Lewontin's term (1978), as we expect populations subject to natural selection to “track” their environment, so should we expect agents so subject better to track their environment by means of improved cognitive equipment—that is, roughly, by means of rationality. Hence it is not surprising that modern progressive naturalists focus on the thin or formal epistemic and practical virtues associated with rationality, a doctrine that might usefully be labeled *evolutionary rationalism*. Strong forms of this doctrine are associated with Quine and Dennett—for example, these well-known epigrams: “Creatures inveterately wrong in their inductions have a pathetic but praise-worthy tendency to die before reproducing their kind” (Quine 1969) and “Natural selection guarantees that most of an organism's beliefs will be true, most of its strategies rational” (Dennett 1987). Ross and LaCasse (1995 , 488) state the central tenet more generally, linking it to the sense of rationality that will be the subject of this chapter: “An immensely complex conspiracy of pressures, some cognitive, some genetic, some that work through social and other environmental selection, drive systems that perpetuate themselves through control and feedback mechanisms to statistically converge towards rationality, in the sense captured by microeconomics.” Beniger 1986 makes a related argument historically, and changes the focus from organisms to organizations. Dennett's early statement (1978 , 16) of his po
end p.420

sition makes the link to game theory explicit: “Game-theoretical predictions applied to human subjects achieve their accuracy in virtue of the evolutionary guarantee that man is well designed as a game player, a special case of rationality” (see Ross 2002). Evolutionary naturalism's general speculative thesis is strongly criticized in Stich 1990 and Kitcher 1985 .
As we shall see in the next section, the convergence of evolutionary and economic game theory allows us to develop evolutionary rationalism in less speculative and more precise terms.

2. Isomorphism

In this section I illustrate the isomorphic relation of evolution to rationality by means of a model that solves a pair of deep problems, one in evolutionary theory and apparently a different one for rational choice. I focus the discussion by limiting rationality to game theory and evolution to evolutionary game theory.

2.1.Sex and Fairness

In most sexually reproducing species, the ratio between males and females is equal. This seems unproblematic—natural—to us, which simply reveals how naive our intuitions can

be. The biological problem is puzzling; generally females are the scarcity constraint on reproduction, so we should expect fewer males in an optimum sex mix. Darwin himself was deeply puzzled by this problem: “I formerly thought that when a tendency to produce the two sexes in equal numbers was advantageous to the species, it would follow from natural selection, but I now see that the whole problem is so intricate that it is safer to leave its solution for the future” (Darwin 1871 , quoted in Skyrms 1996 , 3).
Skyrms (1996 , chap. 1) has us consider Darwin's puzzle in parallel with another: the elementary problem of dividing between two of us a simple good, such as a cake, when neither of us has any special claim. Here the problem is not what we ought to do—sharing equally seems obvious—but *why* we ought to do this. Skyrms argues that rational choice cannot answer this basic question about fairness. Game theory's central concept is the Nash equilibrium: “We have an *equilibrium in informed rational self-interest* if each of our claims are optimal given the other's claim. In other words, given my claim you could not do better by changing yours and given your claim I could do no better by changing mine” (5).

But there is a problem. Each claiming half the cake is a Nash equilibrium, but so is me claiming 2/3 and you claiming 1/3. Skyrms explains, “There is a profusion of strict equilibrium solutions to our problem of dividing the cake, but we want to say that only one of them is *just*. Equilibrium in informed rational self-interestdoes not explain our conception of justice” (5).
Skyrms constructs a model around the simplified game with three strategies for dividing a unit good shown in figure 22.1 . Notice that if the sum of the players' bids is greater than one, then each gets nothing, whereas if the sum is less than or equal to one, each gets their bid.
In figure 22.1 the rows and columns are strategies and the payoffs (to the row strategy in the figure) represent reproductive fitness: the number of offspring expected. Let's think through the dynamics of this situation. If a player modestly demands 1/3, those who demand more will get more and will therefore reproduce more. Similarly, those who are greedy and demand 2/3 will lead to greater numbers of those who demand 1/3. Demand 1/2 “is a stable equilibrium. In a population in which everyone demands half the cake, any mutant who demanded anything different would get less than the population average. Demanding half of the cake is an *evolutionarily stable strategy* in the sense of Maynard Smith and Price, and an attracting dynamical equilibrium of the evolutionary replicator dynamics” (Skyrms 1996 , 11; we return to these concepts in sec. 2.2.1). Thus in the evolutionary model, equal division stands out among the several distributions. In this case evolution is able to solve the equilibrium selection problem “without imposing heroic cognitive requirements on players” (Bicchieri forthcoming).
The evolutionary model can produce a solution where rational choice fails because evolutionary payoffs are in reproductive fitness. In the simplest model of this dynamic process, the replicator dynamics, copies of each strategy are added to the population in proportion to its share of the total payoff. This means that the success of the strategies demand 1/3 and demand 2/3 will be limited by the negative feedback of encountering more copies of themselves. Now we can return to the problem with which we began, to see how evolution explains equal sex ratios. Supposing a parent's inherited tendency toward having more male or female offspring does not effect the number of children the

parent has, Fisher (1930) notes that the strategy “can nevertheless affect the expected number of grandchildren.In a population with a preponderance of males, a genetic tendency to produce more females would spread. There is an evolutionary feedback that tends to stabilize at equal proportions of males and females” (Skyrms 1996 , 7–8).
The explanations based on this model are both stronger and weaker than they might first appear. Returning to the cake division problem, one might think the solution depends on the assumption of an initially equal population mix of strategies. But the argument is stronger, because the evolutionary model generalizes to other initial population mixes. This produces the diagram in figure 22.2 , where the vertices represent populations of pure strategies. (Since the populations must
end p.422

		Demand 1/3	Demand 1/2	Demand 2/3
Modest	Demand 1/3	1/3	1/3	1/3
Fair-minded	Demand 1/2	1/2	1/2	0
Greedy	Demand 2/3	2/3	0	0

Figure 22.1 Dividing the Fitness Cake

add up to one, all possible combinations are restricted to this triangle, the so-called “simplex.”) The flows indicate that most population mixes of the three strategies lead to demand 1/2, except for the lower region, which gets trapped in a polymorphic mix of demand 1/3 and demand 2/3. This last possibility shows how the argument is weaker than one might expect. It involves no claim that equal division is a necessary outcome. Instead the equal outcomes form a basin of attraction the size of which is determined by the structure of the payoffs in the game. In the game under consideration, equality is likely but unequal outcomes are also possible.

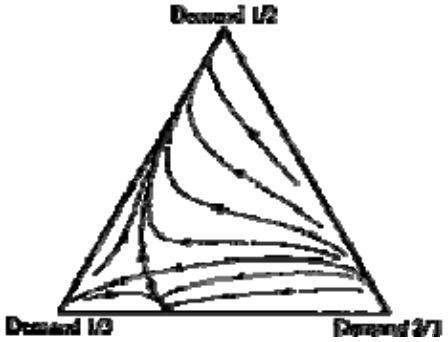

Figure 22.2 Dynamics of Equality. Figure after Skyrms 1996 taken from www.ags.uci.edu/~jalex/latticegmodels. An interactive version of this model is available at www.ethics.ubc.ca/eame/eameweb/Skyrms/chapter 1.htm.

Summing up, the evolutionary model solves both the biological problem about equal sex ratios and the problem concerning the counterintuitive results of rational choice in the divide-the-cake game. Equal sex ratios are an evolutionarily stable strategy in most populations, as are equal demands in the divide-the-cake game. The theoretical foundation of this result is the dynamic evolutionary model
end p.423

that accounts for the two situations. I will not speculate whether this model accounts for our intuitions of fairness but will focus instead on the theoretical foundation in this chapter.

Evolution is isomorphic to rationality only if we restrict the range of both concepts. This section will focus on practical rationality, or deliberation, and in particular I will focus on rationality in interaction. Here is where we would expect a rational and an evolutionary account to differ. Unlike theoretical rationality or practical rationality in a fixed environment, practical rationality where one's interests depend on the actions of others is the most complex problem of rationality. Elster (1984 , 18ff.) distinguishes strategic from parametric (or static) environments: “In the strategic or game-theoretic mode of interaction, each actor has to take account of the intentions of all other actors, including the fact that their intentions are based upon their expectations concerning his own. This was thought for a long time to involve an infinite regress, but frequently this is not the case. Using the concept of an *equilibrium point* it is possible to cut short the infinite regress and arrive at a uniquely definable and predictable course of action that will be chosen by rational men.” Therefore we assume the game theoretic account of strategic rationality, which is the subject of Bicchieri, chap. 10, this volume, and focus on the evolutionary variants of game theory.

2.2.Evolution as Evolutionary Game Theory

The ideas behind evolutionary game theory originate in Fisher 1930 , but the central technique was developed in Maynard Smith and Price 1973 , Maynard Smith and Parker 1976 , and Maynard Smith 1982 . Dawkins 1976 's brilliant exposition, radical formulations, and catchy phrases, like “selfish gene” and “genesmanship,” were very successful at propagating these ideas. Indeed, the second edition (Dawkins 1989) adds an appendix tracking the spread of some of these “memes” in the literature.
Evolutionary game theory differs from economic game theory in several respects. Maynard Smith lists two:
Paradoxically, it has turned out that game theory is more readily applied to biology than to the field of economic behavior for which it was originally designed. There are two reasons for this. First, the theory requires that the values of different outcomes (for example, financial rewards, the risks of death and the pleasures of a clear conscience) be measured on a single scale. In human application, this measure is provided by “utility”—a somewhat artificial and uncomfortable concept: in biology, Darwinian fitness provides a natural and genuinely one-dimensional scale.Secondly, and more importantly, in seeking the solution of a game, the concept of human rationality is replaced by that of evolutionary stability. The advantage here is that there are good theo
end p.424

retical reasons to expect populations to evolve to stable states, whereas there are grounds for doubting whether human beings always behave rationally. (Maynard Smith 1982 , vii)
Third, in an amusing contribution, Krugman (1996 , 2) claims, “The main difference between evolutionary theory and economics is that while economists routinely suppose that the agents in their models are very smart about finding the best strategy—an economist is always defensive about any model in which agents are assumed to act with less than perfect rationality—evolutionists have no qualms about assuming myopic behavior. Indeed myopia is of the essence of their view.”
Evolution is one alternative to rationality, but we should avoid the temptation to make evolution the sole general purpose alternative to rational choice. In addition to evolutionary selection, agents may learn and be influenced by social norms. The application of evolutionary game theory to these latter cases is less well established. Historically, evolutionary game theory was introduced into biology, but the theory has been generalized on the one hand and limited to human agents on the other. I note three variants of evolutionary game theory.

2.2.1. Biological Evolutionary Game Theory

The central notion in biological evolutionary game theory is an *evolutionarily stable strategy* (ESS) introduced by Maynard Smith and Price attempting to explain why most intraspecific animal conflicts are “limited war” rather than “total war” types: “The concept of an ESS is fundamental to our argument; it has been derived in part from the theory of games, and in part from the work of MacArthur and of Hamilton on the evolution of the sex ratio. Roughly, an ESS is a strategy such that, if most of the members

of a population adopt it, there is no `mutant' strategy that would give higher reproductive fitness" (1973, 15).
Formalizing this definition links the ESS with game theory's Nash equilibrium concept (see Bicchieri, chap. 10, this volume). Strategy x is an ESS if, for every strategy y distinct from x and utility function u (where $u(x,y)$ is the utility of "playing" x against y),

1. $u(x,x)$;eg $u(y,x)$
2. If $u(x,x) = u(y,x)$ then $u(x,y) > u(y,y)$

According to Samuelson (1997 , 40), "This, the original, definition of an ESS offered by Maynard Smithhas the advantage of making it clear that the ESS condition is the combination of a Nash equilibrium requirement, given by (1) and a stability requirement, given by (2). The latter ensures that the ESS [x] can repel mutants."
In addition to an account of equilibrium selection, evolutionary game theory "provides an explanation of the dynamics of the selection process, something
end p.425

which [game theory's] refinement program cannot do" (Bicchieri forthcoming).
Returning to the simplest case, the replicator dynamics, "every ESS is a strongly dynamically stable (or attracting) equilibrium in the replicator dynamics" (Skyrms 2000 , 274). As Samuelson notes, "The best-known form of the replicator dynamics was introduced by Taylor and Jonker 1978 and by Zeeman 1992 but similar ideas have been used in earlier applications. Hofbauer and Sigmund [1988] discuss one of the earliest analysesin which the mathematician V. Volterra constructed a dynamic model to explain population fluctuations of predator and prey fish species in the Adriatic Sea" (1997, 18).
Strictly speaking, the replicator dynamics models a simplified *ecological approach* to dynamics, as the types of interactor in the population are fixed. Axelrod (1981) distinguishes an *evolutionary approach* where new types of interactor can enter the population. He later pioneered using the genetic algorithm to implement this more complex model in Axelrod 1987 . (Axelrod's work is discussed further in sec. 4.1 below.)
The genetic algorithm, pioneered by Holland (1975), mimics the mechanism of biological evolution to generate new variants as well as to test them by differential selection. Typically pairs of successful functions, composed of vectors of bits, are combined by crossing over part of each parent vector; vectors are also altered by point mutation. The genetic algorithm allows a model to explore a much larger space of agent interactions than is possible by enumeration of agent types combined with the replicator dynamics.

2.2.2. Economic Evolutionary Game Theory

Where biological evolutionary game theory is intentionally broad in the scope of its agents, economic evolutionary game theory focuses more narrowly on explaining human action. Economic evolutionary game theory can be seen as a reply to Maynard Smith's point quoted above about the replicator dynamics applying better to nonhuman organisms. Mailath (1992 , 268) agrees: “There is nothing in economics to justify the replicator dynamic.” Therefore, economic evolutionary game theory modeling is based on a human learning dynamic. The “model consists of a large population of myopic and unsophisticatedplayers.[An unsophisticated player] does not conduct his calculations under the assumption that the other players are similarly reoptimizing.[The players] can observe the history of plays” (Mailath 1992 , 261). Samuelson 1997 makes a similar distinction between economic and biological evolutionary game theory.

2.3. Evolutionary Generalism

At the other extreme, Skyrms is an “evolutionary generalist” (D'Arms, Batterman, and Gorny 1998) who aims to model evolutionary processes in the most abstract form. He explicitly states that his models are sufficiently general to cover both biological and cultural evolution: “[Fair division's] strong stability properties guarantee that it is an attracting equilibrium in the replicator dynamics, but also make the details of the dynamics unimportant. Fair division will be stable in any dynamics with a tendency to increase the proportion (or probability) of strategies with greater payoffs.For this reason, the Darwinian story can be transposed into the context of cultural evolution, in which imitation and learning may play an important role in the dynamics” (Skyrms 1996 , 11). Similarly, Dawkins 1983 and Dennett 1978 , chap. 5, make claims for the universality of Darwinism.

Note that when we generalize evolutionary models to include social as well as biological evolution, we can no longer always rely on Maynard Smith's objective Darwinian alternative to subjective utility functions. This is because in social evolution reproductive success can be mediated by agents who choose which agents' strategies to imitate. Indeed, since these copiers need to be able to compare different agents' utilities, they require a more demanding subjective utility function (Bicchieri forthcoming).

2.4. Why Isomorphism Is Important

To return to our opening quotation, Skyrms draws an important conclusion from the way game theory shows evolution and rationality to be isomorphic:

The most striking fact about the relationship between evolutionary game theory and economic game theory is that, at the most basic level, a theory built of hyper-rational actors and a theory built of possibly non-rational actors are in fundamental agreement. This fact has been widely noticed, and its importance can hardly be overestimated. Criticism of game theory based on the failure of rationality assumptions must be reconsidered from the viewpoint of adaptive processes. There are many roads to the Nash

equilibrium concept, only one of which is based on highly idealized rationality assumptions. (Skyrms 2000 , 273)

3. Differences

Having established the strong similarity of evolution and rationality, their differences become interesting. Elster 1989 , chap. 8; Skyrms 1996 ; Skyrms 2000 ; and Sober 1998 all survey differences between rationality and evolution. To return
end p.427

again to our opening quote, Skyrms continues: “However,the situation ismore complicated than it might at first appear. There are aspects of accord between evolutionary game theory and rational game theory as well as areas of difference. This is as true for social evolution as for biological evolution. The phenomena in question thus have considerable interest for social and political philosophy.” (Skyrms 2000 , 273). “Throughout a range of problems associated with the social contract, the shift from the perspective of rational choice theory to that of evolutionary dynamics makes a radical difference. In many cases, anomalies are explained and supposed paradoxes disappear” (Skyrms 1996 , xi). We will consider three differences.

3.1. Symmetry

“A single population evolutionary setting imposes a symmetry requirement which selects Nash equilibria which appear implausible in other settings” (Skyrms 2000 , 273). Skyrms (1996 , x) “shows how evolution imposes a `Darwinian veil of ignorance' that often (but not always) leads to selection of fair division in a simple bargaining game. In contrast, rational decision theory leads to an infinite number of equilibria.” This case was discussed in section 2.1 above.

3.2. Dominated Strategies

“Refinements of the Nash equilibrium are handled differently. Standard evolutionary dynamics, the replicator dynamics, does not guarantee elimination of weakly dominated strategies. In a closely related phenomenon, when we consider extensive form games, evolutionary dynamics need not eliminate strategies which fail the test of sequential rationality” (Skyrms 2000 , 273). In a chapter on “Commitment,” Skyrms (1996) uses the asymmetrical game of Take It or Leave It to demonstrate these differences. This game puts player 1 in a privileged position, able to make player 2 a fair (demand 5) or an unfair (demand 9) offer, which the latter must then accept or reject (getting nothing). In this

sequential ultimatum game, (see figure 22.3), human experimental subjects moving first often make fair initial offers, while the rational strategy is to offer the second player the smallest possible share—the solid lines, leading to the {9,1} payoff. Various explanations of these results do not go very deep. If humans prefer fair outcomes, why do they have these preferences? “Appeal to norms of fairnesshardly constitutes an explanation in itself.Why have norms of fairness not been eliminated by the process of evolution?” (Skyrms 1996 , 28).
Skyrms constructs an evolutionary model of the ultimatum game. This model
end p.428

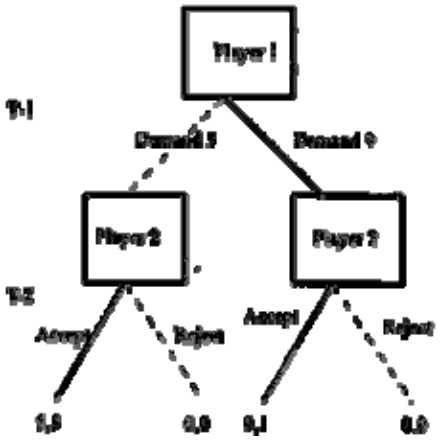

Figure 22.3 Ultimate Game.

is more complex than that discussed in section 2.1, since each player needs, when playing the role of player 2, to be able to react to the two strategies open to player 1.Thus there are eight possible strategies, including Gamesman, who demands 9 and accepts all offers, Fairman, who demands 5 and accepts 5 but rejects 9, and the counterintuitive Mad Dog, who demands 9 but rejects 9 and accepts 5. The first two are the strategies that interest us initially, as the first is the recommendation of rational choice theory, and the second, although it would explain the experimental results, is rationally defective. By rejecting demands of 9 (offers of 1), Fairman threatens to choose what is worse (0 over 1) and thus fails the test of “modular rationality,” which requires that a strategy “specify a *rational* choice at each choice point” (Skyrms 1996 , 24).
Skyrms shows that although the Fairman strategy is dominated by Gamesman—the latter never does worse and sometimes better—some initial populations will allow Fairman to persist (see figure 22.4). This is a case where misleading results can follow from oversimplified evolutionary models. Skyrms points out that, because it fails to generate the strategies that would do better, the replicator dynamics can leave dominated strategies in place. In this way, the replicator dynamics are conservative. He shows that introducing crossover (with a genetic algorithm; see above, sec. 2.2.1) need not correct this bias; mutation is needed to test robustness fully. Thus he stresses the important differences, with respect to contrasting evolution and rational choice, between variations of the genetic algorithm that are often treated as merely technical variants. Nonetheless, even

the more refined evolutionary model will allow dominated strategies to persist. This sounds like good news for Gauthier (1986) and McClennen (1990), who have defended dominated strategies of commitment as the doctrines of constrained maximization and resolute choice, respectively. Perhaps evolutionary game theory can add new life to these theories, which have not been well received as contributions to economic game theory; see, for example, Binmore 1994 . We look at one way this might play out in the concluding section.
end p.429

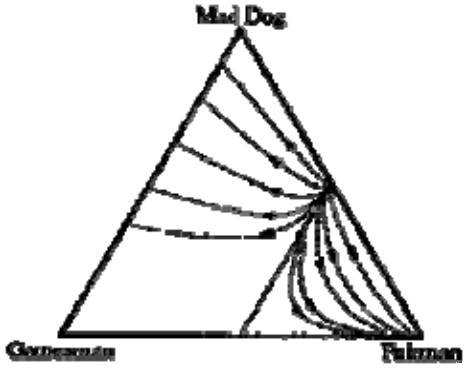

Figure 22.4 Dynamics of Ultimatum Game. Figure after Skyrms 1996 taken from www.ags.uci.edu/~jalex/latticegmodels. An interactive version of this model is available at www.ethics.ubc.ca/eame/eameweb/Skyrms/chapter 1.htm.

Samuelson (1997 , 34) suggests that these results will have general impact on the theory of rational choice: “The failure to eliminate weakly dominated strategies is characteristic of a wide variety of evolutionary models. I suspect that if there is any general implication to emerge from evolutionary game theory, it will be that we should be much less eager to discard weakly dominated strategies and should place less emphasis on dominance arguments in selecting equilibria. This in turn may lead to reconsideration of applied work in a wide variety of areas.”

3.3. Actual Fitness and Intended Outcomes

A third difference between rationality and evolution has to do with the payoffs. “Rational choice is concerned with the intended outcomes of action. Selection mechanisms operate through actual outcomes. In explanations of animal behavior, where intentions have at best a minimal place, actual outcomes must bear most of the explanatory burden. It is more controversial which mechanism is the most important in the study of human action” (Elster 1989 , 71).
Binmore (1998a , 33) appeals to this difference to explain the previous result: “The evolutionary argument holds together `tolerably well' when used to defend the notion of a

Nash equilibrium but the abstract form of the argument comes under strain when applied to subgame-perfect equilibria because this notion depends on postulating the players would behave consistently even under circumstances that never actually occur. The orthodox neoclassical notion of consistency then parts company with the evolutionary story since the irrational plans can only be selected against if they are sometimes brought into play."
This third difference lets us correct what may be a misunderstanding of the impact of evolutionary game theory literature. Starting with Axelrod (1984), there has been a tendency of some to see evolutionary models as generally supporting
end p.430

cooperation in cases—like the Prisoners' Dilemma—where rationality would not. However, as Sober (1998 , sec. 4) points out, the third theoretical difference allows that evolutionary selection will sometimes fail to cooperate where rational agents, looking ahead to the consequences of their own defection, would cooperate. Generalizing this point, I have argued elsewhere (Danielson 1998a) that one should not identify evolutionary models with assumptions about positive correlation of similar agents, which can be used to derive more cooperative outcomes.

4. Simulations and Multilevel Models

Roughly, the contrast between rationality and evolution projects onto methods, with evolution characterized more by simulations and rationality by more formal models. We begin with Axelrod's influential work and the criticism it has attracted, move on to general skepticism about evolutionary simulations, and end more constructively with some recent work on multilevel simulations. Incidentally, as Dawkins (1976 , chap. 4) points out, the use of simulations is related to our topic in a more general way. For the same reasons of efficiency that we researchers build computer simulations to predict what might happen under various assumptions, so too evolution is drawn to produce agents that model their environments—that is, to become rational, responding to indirect reinforcements, and in Popper's famous phrase, "let [their] hypotheses die in [their] stead" (quoted in Dennett 1995 , 375).

4.1. Axelrod

While evolutionary biology provided one path into contemporary evolutionary game theory, the political scientist Robert Axelrod's tournament experiments and simulations (1981 , 1984 , 1987 , 1997) provided another. Axelrod invited the research community to submit strategies (programmed in FORTRAN) for an indefinitely repeated Prisoners' Dilemma, which he played in a round-robin tournament. He then applied the replicator dynamics to simulate further rounds of the tournament. Critics of Axelrod lament two

aspects of his influential work: first, that he simulates what should be obvious from theory—that is, that there are infinitely many equilibria in the repeated Prisoners' Dilemma (the so called folk theorem). Second, that he overemphasizes one equilibrium strategy: tit-for-tat. The second is an important criticism, but the first is unfair. Axelrod introduced important new methods and initiated an interdisciplinary dialogue based on them. Notice, for example, that Axelrod (1987) initiated the use of programmed strategies, which was influential in computer science, as well as the use of genetic algorithms to generate strategies, which brought a previously disconnected field into closer contact with economics and biology. Unfair, because some of the best results refining Axelrod's work do not derive from the pure theory that he "should have" deployed, but by means of the new methods that he pioneered. For example, Binmore and Samuelson (1992) build populations of finite state machines.
Binmore (1998b) assesses Axelrod's contribution:
The *folk theorem* of game theory proved by several authors simultaneously in the early fiftiesdescribes in precise detail *all* of the outcomes of a repeated game that can be sustained as equilibria. However, although Axelrod didn't discover the folk theorem, I believe that he did make an important contribution to game theory, which has nothing to do with the particular strategy TIT-FOR-TAT nor with the mechanisms that *sustain* any of the other equilibria in the indefinitely repeated Prisoners' Dilemma. He did us the service of focusing our attention on the importance of evolution in *selecting* an equilibrium from the infinitude of possibilities whose existence is demonstrated by the folk theorem. Other game theorists may protest at my recognizing someone who knew no game theory at the time he made his contribution and still resolutely ignores game-theoretic commentary on his work, but it is inescapable that the evolutionary ideas pioneered by Axelrod now provide the standard approach to the equilibrium selection problem in game theory.

4.2. A Methodology: Multilevel Modeling

So far we have emphasized the similarities and differences between evolution and rationality. Stepping back, it is obvious that the complete picture must include both; evolution has historically produced some rational agents. So the question arises, how might one model the interaction between evolution and rationality? Given the difference in time scales of the two processes, a natural way to approach the relation of rationality and evolution is by a two-level model. Binmore (1994 , 151) suggests a framework:
The principal agent problem involves a principal with certain aims who can only act through agents who may have aims of their own. The principal therefore seeks to design an incentive scheme that minimizes the distortions resulting from having to delegate to the agents. Nature, as principal, is blind, but her agent, homo economicus, can see his environment. Thus, although Nature loses fine control being unable to modify his environment directly, she gains access to information that would not otherwise have been available to her. If the environment changes sufficiently rapidly, the gains will outweigh the losses.
end p.432

This framework suggests using an evolutionary mechanism to generate and select agents' preferences while a rationality mechanism uses those preferences to select strategies. Levine (1998), Sethi and Somanathan (2001), and Danielson (2002) explore questions of reciprocity and fairness employing models of this form; Harms (1997) applies a two-level model to questions in evolutionary epistemology.

Of course, simulation is not limited to *two* levels of modeling, although the more complex the simulation the more difficult it is to understand its output. Epstein and Axtell 1996 is an innovative exploration of a multilevel simulation, incorporating not only evolution of agents who make (boundedly) rational choices (of movements) but also evolution of their immune systems and social identity tags.

4.3. Computational Differences

We have focused on comparing rationality and evolution rather narrowly as formal accounts of practical rationality in social interaction, abstracting away most computational problems by using equilibrium concepts. But if we broaden our contrast, rationality looks weaker. In their provocatively titled "Better than Rational: Evolutionary Psychology and the Invisible Hand," Cosmides and Tooby (1994) argue,

One point is particularly important for economists to appreciate: it can be demonstrated that "rational" decision-making methodsare computationally very weak: incapable of solving the natural adaptive problems our ancestors had to solve reliably in order to reproduce.This poor performance on most natural problems is the primary reason why problem-solving specializations were favored by natural selection over general purpose problem-solvers. Despite widespread claims to the contrary, the human mind is not worse than rational (e.g. because of processing constraints)—but may often be better than rational. On evolutionarily recurrent computational tasks, such as object recognition, grammar acquisition, or speech comprehension the human mind greatly outperforms the best artificial problem-solving systems that decades of research have produced, and it solves large classes of problems that even now no human-engineered system can solve at all.

However, we should not draw too practical a conclusion from this claim. It doesn't follow that one can better engineer a general purpose robot or a chess-playing automaton by applying evolutionary rather than rational techniques. The competition between these methodologies in artificial intelligence is a matter of lively controversy, ranging from strong claims for evolutionary techniques (Koza 1992 ; Pfeifer and Scheier 1999) to this dry "logicist" reply in a leading AI textbook: "Like neural networks, genetic algorithms are easy to apply to a wide range of

problems. The results can be very good on some problems, and rather poor on others.Denker's remark `neural networks are the second best way of doing just about

anything' has been extended with `and genetic algorithms are the third'" (Russell and Norvig 1995 , 621).

5. Normativity

Although we began with a discussion of the normative limitations of appeals to evolution, so far we have considered evolution and rationality as mechanisms without regard to their normative or descriptive use. I turn to this issue in this section and close with some intriguing conjectures about the effect of theory on the evolution of human rationality.

5.1. Normative Rationality versus Descriptive Evolution

As other chapters in this volume discuss the normative status of rationality, here we focus on how having an alternative evolutionary mechanism influences this issue. We have already noted in section 1 that the isomorphism between evolution and rationality affects their role in normative arguments. Rational choice only transfers value to alternatives valued by preferences; evolution optimization matters only if one values fitness. Critics of normative appeals to evolution were typically aiming at more substantial theses, such as the claim that complexity is good because organic complexity is the goal of evolution. The standard view is that while rational choice is normative, evolution is merely descriptive. Nonetheless, the connections between rationality and evolution do give the latter a new normative significance.

First, there is the philosophical question why rationality but not evolution should be seen as normative, given their strong formal similarity. One can, of course, criticize evolutionary mechanisms for not selecting a global optimum when faced with a local optimum. But conversely, one can criticize the rationality mechanism for limiting an agent to the alternatives she conceives of. Similarly, the driver of evolutionary models is reproductive fitness, the value of which is open to the antinaturalistic challenge "Why should I care about being copied?" (Or more generally, what de Sousa (1980) calls Oscar Wilde's Boomerang: "The first duty in life is to be as artificial as possible.") But the analogous rejoinder

here, leading to "Why should I care about my own preferences?", is only disallowed because preferences are defined as determining behavior. If we consider a more complex case, where an agent finds herself choosing to smoke or drink against her acknowledged interest in health or social life, matters are less clear (Varner 1998 , Rachlin 2000).

Second, recent normative theory gives rationality a greater normative role and this gives evolution a new role as well. Starting with Rawls's (1971) influential contractarian scheme to make the theory of justice part of the theory of rational choice, Gauthier (1986) and Binmore (1994 , 1998a) develop and refine the role of rationality in moral theory.

Therefore, when evolution modifies the theory of rational choice, it has normative consequences as well. For example, Skyrms (1996 , 6) argues that the indeterminacies of rational choice also infect normative theories built upon it. Following his argument that straight rational choice fails to find a solution to the simple bargaining problem (discussed in sec. 2.1), he claims that the social contract theories of Rawls 1971 and Harsanyi 1955 , which invoke the device of the veil of ignorance, behind which neither of us knows which piece of cake he will get, also fail to account for the salience of the equal outcome.

Third, as we have seen, evolutionary arguments may change our conception of rationality. Samuelson (1997 , 3) notes the philosophical uncertainty of key rationality concepts: “For some, game theory is a normative exercise investigating how decisions *should* be made.The intricacy of this problem grows out of the self-reference that is built into the notion of rationality, with all of its attendant logical paradoxes. The richness of the problem grows out of our having no a priori notion of what it means to be rational, substituting instead a mixture of intuition, analogy and ideology. The resulting models have a solid claim to being philosophy and art as well as science.” Evolutionary equilibrium selection may influence these intuitions. For example, “the restrictions to be placed on out-of-equilibrium behavior may appear to be obvious in the Chain-Store Game, but more formidable challenges for an equilibrium refinement are easily constructed, involvingmore intricate challenges to our intuition as to what is a `good' equilibrium. [This literature] has produced an ever-growing collection of contending refinements, many of which are quite sensitive to various fine details in the constructions of the model, butvery little basis for interpreting these refinements or choosing between them” (Samuelson 1997 , 9).

Fourth, while evolutionary game theorists have been content to see the theory as “a positive exercise in investigating how decisions *are* made” (Samuelson 1997 , 4), we should note the interdependence of positive and normative considerations. In contexts so dependent on mutual expectations, positive theory has great normative influence. If I have strong reasons, based on evolutionary equilibrium selection, to expect you to choose a particular strategy, normative rationality gives me a reason to coordinate with your choice.

end p.435

More generally, Skyrms (1996 , 108–9) sees a broadly normative role for evolutionary game theory:

What does this all have to do with ethics and political philosophy?I have not said anything about how human beings should live their lives or how society should be organized.There is, nevertheless, a conception of these fields under which this book falls. Ethics is a study of possibilities of how one might live. Political philosophy is the study of how societies might be organized. If possibility is construed generously we have utopian theory. Those who would deal with “men as they are” need to work with a more restrictive sense of possibility. Concern with the interactive dynamics of biological evolution, cultural evolution, and learning provides some interesting constraintsEven those who aim to change the world had better first learn how to describe it.

5.2. What Are the Prospects for Convergence?

We have emphasized how recent is our appreciation of the strong connection between rationality and evolution. One difference, of course, is that evolution began as a theory of mindless organisms while game theoretic rationality began as an account of maximally reflective agents. So we close with a pair of contrasting speculations about the effects of our new theoretical insights into the further evolution of human rationality. Both predict that social outcomes will depend on the theory widely held by agents, and this theory will, in part, be selected by social evolution. They differ on the theory recommended.
Binmore favors game theory as a normative theory. “If such a theory is ever to do more than grace the dusty shelves of academic libraries, it will be because the combination of education and evolution drives society in the direction of the theory. In this process, the existence of the theory itself will be an important factor. A widely applicable theory of games would, of necessity, involve a strong element of self-prophecy in the sense that the existence of the theory itself would be partly responsible for bringing about stabilizing the event which it `predicts'” (1990, 18–19).
McClennen is a critic of the game theoretic account of rationality, which he portrays as an “extremely limiting sort of model” that begs many important questions. So he suggests an evolutionary thought experiment, in which individuals are informed of the possible advantages of more complex forms of decision making, including principle-constrained choice in cooperative dilemmas. (In sec. 3.2 I mentioned how McClennen's proposal might be selected in an evolutionary model using the Ultimatum Game.) “One might then reasonably expect to see this more efficient mode of dynamic decision-making drive out more costly precommitment and enforcement methods, and this through nothing more than what economists like to describe as *the ordinary competitive process*.” (Danielson 1998b , quoting an earlier version of McClennen 1998).

NOTES

Sections 2.1 and 2.3 draw on Danielson 1998a .

1. Note that when Sober wrote this in the mid-1980s the tragedy of the commons was still evidently better known than the Prisoners' Dilemma model, which Sober introduces in a footnote.

END