Practical Reason and Preference

CHRISTOPHER W. MORRIS AND ARTHUR RIPSTEIN

The traditional theory of rational choice begins with a series of simple and compelling ideas. One acts rationally insofar as one acts effectively to achieve one's ends given one's beliefs. In order to do so, those ends and beliefs must satisfy certain simple and intuitively plausible conditions: For instance, the rational agent's ends must be ordered in a ranking that is both complete and transitive, and his or her beliefs must assign probabilities to states of affairs relevant to the achievement of those ends. The requirement of completeness ensures that all alternatives will be comparable; the transitivity condition ensures that at least one alternative will be ranked ahead of the others in each situation. If the completeness condition is violated, the agent will not always be able to compare alternatives and consequently to make a choice. If transitivity is violated, a situation may arise in which the agent will be unable to achieve his or her ends because for any alternative there will be another that will be preferred to it. On the belief side, there are similar requirements of completeness and consistency: An incomplete ordering of beliefs might recommend no action at all, and inconsistent beliefs might recommend incompatible courses of action. Provided that the constraints are satisfied, whenever the opportunity to make a decision presents itself, the rational agent will choose the course of action that will be most likely to achieve his or her ends. Preferences and beliefs both enter into that evaluation: A highly valued outcome whose probability of achievement is low may be ranked lower than a less valued outcome that enjoys a higher probability of success. Once preferences and probabilities are fixed, the rational agent acts in the way that will directly maximize his or her expected payoff.

The theory treats reason as an instrument for achieving one's ends, whatever those ends may be, or so it is usually interpreted. Formal treatments often work with monetary examples, sometimes creating the impression that the theory is committed to some conception of self-interest or, worse, that it must put monetary values on alternatives. But although the interests with which the theory works are always the interests *of* a self, they need not be interest *in* oneself. Persons concerned to promote the interests of others or to set up just and equitable social institutions also have reasons to pursue those

ends based on their estimates of their importance and the probabilities for various means of achieving them.

The forward-looking nature of the theory requires that it judge possible courses of action on the basis of their expected consequences. Past actions may provide grounds for expectations about what others will do in the future, but they cannot provide grounds for preferring one alternative to another, as the past cannot be changed. Importantly, this forward-looking focus seems to require one to treat all commitments as irrational or empty. Commitments are backward-looking and also require one to make decisions on the basis of general rules or principles, thus ignoring the specific consequences of particular acts. From the standpoint of the traditional theory, any such commitments either will be redundant, recommending the same course of action that forward-looking rationality recommends, or else will be contrary to reason, recommending a different one.

The hostility of the traditional theory to commitment gives rise to two distinct sets of problems. One set is internal to the theory itself: In a variety of familiar situations involving a plurality of interacting persons, the direct pursuit of one's ends makes cooperation impossible and thus seems in some sense self-defeating. The best-known example is the infamous prisoner's dilemma. In the original tale, two prisoners have been caught committing a crime, and each faces a short jail term, but each is offered a chance to further reduce his or her sentence by testifying against the other. If both remain silent, both will face the short jail term, but each can see that whatever the other does, he or she will do better by confessing. If the other remains silent, confessing will lead to no prison sentence; if the other confesses, doing the same will be the best option. Each thus has an incentive to confess, and the result will be that if both confess, both will spend more time in prison. The example supposes that each is concerned only to minimize his or her own prison time, but the underlying problem is structural, rather than one of selfishness as such. For example, parents concerned to provide for their children, or to protect them from automobile accidents, may find themselves facing a similar structure of incentives. A number of parents might choose to buy larger and larger cars in anticipation of other parents, equally concerned for the welfare of their children, driving larger cars. Each can see the advantages of driving a larger car, whatever others do, so all choose to drive larger cars, thus providing their children with protection from an increasing peril that is itself simply the result of each parent's attempt to protect his or her children. All would be better off if all could commit to small cars. But none can commit. Economic exchange often has the structure, at least on the surface, of a prisoner's dilemma.

The obvious solution in these situations is for agents to make some sort of agreement in anticipation of the situation. The difficulty, of course, is that the

keeping of the agreement is subject to the same incentives, and so although all parties can see the clear advantages of making such an agreement, they will have incentives to defect from it, and consequently the agreement will be unstable. Thus all will end up worse off because of their inability to honour commitments or to abide by mutually beneficial agreements. Another possible solution for similar problems is the recognition of an authoritative decision-maker, someone whose directives are reasons for action and who could thus enable agents to cooperate on mutually beneficial outcomes. But again the problem reappears: Agents have incentives to disregard the allegedly authoritative directives and directly to select the individually most advantageous course of action.

The other obvious solution is coercive: to empower some person or agency to intervene so as to change the incentive structure (for instance, by imposing sanctions for defection), so that all will act in ways that will not be collectively self-defeating after all. In the context of the prisoner's dilemma, members of a criminal group might execute those of their number who testify for the state; in the parent's dilemma, the escalation might be halted by taxing larger vehicles. The difficulty with the coercive solution is not that it is impossible – as the examples suggest, such solutions are used all the time – but that it is wasteful, because it requires the expenditure of additional resources to get people to do what it is in their own interest to do anyway.

The second problem is related, but is, strictly speaking, external to the theory. This time the problem is that the making and keeping of commitments appear to be rational processes, not, perhaps, in the sense that is required by the traditional theory, but in the sense that it makes sense to stand back from particular agreements and evaluate them. In our earlier examples, the commitments that proved impossible to keep involved cooperation with others, but the external problem arises for a broader range of commitments. It is not uncommon to have the thought that one should not have agreed to something, or should have sought more favourable terms for something to which one did agree. Nor is it uncommon to regret one's failure to stick by a commitment one has made, even if one does not doubt that acting contrary to it was, at the time, the best way to secure one's ends. The straightforward or direct maximizer of orthodox rational choice theory cannot have such thoughts, for such an agent does not regard any commitments as binding. Yet the familiarity of such thoughts, and the way in which they are disciplined by considerations about the consequences of competing courses of action, should give us pause, for the traditional theory treats all commitments as alike, and none of them as rational, in ways that ordinary reasoning about them suggests they are not.

Among his many contributions to moral and political philosophy, David Gauthier has made two signal contributions to the debate about practical rationality and to the concerns we have raised. First, he has clarified the

concept of a preference, thus giving further specificity to the idea that preferences should be well ordered. The traditional theory of rational choice developed both as an explanatory theory and as a normative theory. The explanatory theory, at the heart of neoclassical economics, showed that the assumption of individual rationality could be used to model a wide range of economic behaviours. By viewing agents as maximizing their own utility, their behaviours in markets and the behaviour of those markets can be rendered orderly and, to some extent, predictable. Such modelling makes it possible to read a person's preferences off of his or her choice behaviour. As a normative theory, rational choice involves advising people about their best courses of action, given their beliefs and preferences. The two roles for a single theory lead to a tension between them: The explanatory model reads preferences off of behaviour by assuming that agents are rational, while the normative theory takes preferences as its starting point and asks what reason demands, without supposing that the agent will live up to those demands. To render them consistent, preferences must be defined independently of choice behaviour, but must remain subjective in an appropriate sense. For Gauthier, to behave rationally is not just to do what one is disposed to do. Nor is it to possess a well-ordered set of preferences. Instead, rational action maximizes the expected satisfaction of expressed and revealed preferences. To the extent that these two dimensions of preference diverge, rationality cannot give consistent advice. Gauthier's key idea is that reason demands the satisfaction of preferences for which attitudinal and behavioural dispositions converge.

Gauthier's revisionist account of the nature of preference provides the first piece in his well-known account of the rationality of commitment. Separating rationality from choice behaviour makes room for a distinction between cases in which an agent has behaved contrary to reason and those in which his rationality has remained intact but his preferences have changed. Reason is demanding enough to require a choice contrary to the disposition to choose differently. Rational commitment requires that reason be unyielding in the face of temptation.

Gauthier's central contribution to the theory of rational choice comes with his account of the rationality of commitment. The central idea is straightforward and elegant: The formal theory of rational choice can operate not only on the choice of action but also on the consideration of dispositions or principles of action. The rational agent thus has reasons, traceable ultimately to his or her own preferences, to adopt a disposition or principle to behave cooperatively and to keep the commitments he or she has reason to undertake. The agent has those reasons because, as the prisoner's dilemma illustrates, the ability to undertake and to keep commitments has advantages in terms of whatever ends an agent might have. The advantages are available only to beings who are able to disregard competing incentives and so to honour the

commitments they have undertaken. The rational agent can appreciate the advantages of being that kind of agent and, in light of them, decide to adopt the appropriate disposition or principle. The decision to adopt a particular disposition affects a broader range of future behaviours than, for example, the decision whether or not to confess to a crime or the decision to purchase a particular automobile. But the choice is forced in the same way that other decisions are: Once a rational agent is aware of the structures of interaction in which commitment is advantageous, that agent cannot decide to forgo the choice of dispositions or principles. To do so would be irrational, for it would be to adopt a course of action with a lower expected payoff than another course that is available.

At the same time, as the dilemmas of parents and prisoners also show, the disposition to cooperate is not always advantageous. It is only when cooperating with others who have similar dispositions or principles that advantage can be had. Otherwise, those who keep agreements open themselves up to exploitation. Gauthier's solution to this problem rests on an idea that is familiar from ordinary life, though difficult to reconstruct in the traditional vocabulary of rational choice: People are capable of discriminating between fellow cooperators and potential exploiters, and they are also quite good at deciding which other people to trust.[1] (Con-artists provide only an apparent exception. They often can exploit trusting people, but their opportunities arise only because most people are trustworthy, and so most people do well by being trusting.) Indeed, Gauthier's account shows the rationality of developing and exercising those discriminating abilities. The capacity to distinguish cooperators from exploiters is useful to the extent that one is capable of entering into mutually advantageous cooperative relations with others; mutually advantageous cooperation is possible for rational agents only if they have that ability. It seems, then, that people can make and keep commitments without external incentives to compliance. The question is how to understand this capacity.

This disposition to cooperate is not a generic disposition to commit to any course of action that offers some advantages as against non-cooperation. Instead, according to Gauthier, it is a narrower disposition to cooperate only on terms that are fair. The basic idea is, again, simple and elegant: The rational agent choosing which disposition or principle to adopt will enter into all and only those cooperative arrangements that will yield fair and mutually advantageous outcomes. Any broader disposition would open one up to exploitation by those with narrower dispositions, and any narrower disposition would turn one into an identifiable exploiter, and so preclude participation in advantageous cooperation.

The status of Gauthier's project is analogous to that of the Hobbesian project of showing the advantages of the authority or sovereignty of states.

Gauthier's solution is internal and dispositional, rather than external and coercive. But the two accounts are alike in their emphasis on the rationality of submitting one's conduct to mutually advantageous norms. There is also an analogy to the Hobbesian problem of leaving "the state of nature": There is generally no point to constraining one's actions by a cooperative disposition or principle unless others do so as well. Hobbes's solution to this problem about leaving the state of nature was to argue that the case for "instituting" a sovereign state carries over to what he called "sovereignty by acquisition." Gauthier's analogous solution is to offer a rational reconstruction of what we might call "morals by socialization" as morals by agreement. It is a reconstruction in the sense that it is meant to explain the rationality of abilities that ordinary people plainly have, rather than being an apologia that must somehow suffice to convince any rational agent to behave morally. Indeed, his solution is not applicable to all rational agents: Opaque beings would not derive the full benefits of constrained maximization, because they would hesitate to depend on each other. And a single opaque agent such as Gyges, the Lydian shepherd of Book II of Plato's *Republic*, might benefit by making commitments but would gain nothing by being able to keep them. For cooperation to be rational, agents must be at least partially transparent – "translucent," in Gauthier's phrase – to each other.

If the revised account of rationality – constrained maximization or, to borrow Edward McClennen's term, "resolute choice" – explains the capacity for commitment presupposed by morality, its theoretical and practical interest is far broader. In particular, threats require the same kind of commitment as morality does, but they need not make cooperation or fair terms their subject matter. In order for a threat to induce others to behave as one wishes, it must be credible. Yet carrying out threats is usually costly, and once a threat has failed to induce the desired behaviour, there is, from the standpoint of the straightforward maximizer, no further point in carrying it out, because no further benefit can be expected from doing so. (Assuming, for the moment, that punishing in this case does not provide a significant signal about one's practice in future cases.) But just as this course of reasoning is available to the threatener, so too is it available to the person threatened, who may conclude that the threat is empty. Thus the sole concern with the future leaves the straightforward maximizer incapable of threatening convincingly. Constrained maximization or resolute choice provides a way of explaining how threats can be rational, and thus how they can be possible.[2]

Gauthier's account thus responds to the two objections to the traditional theory of rational choice. It addresses the internal objection by showing that commitment is rationally defensible in the situations in which it is advantageous. It addresses the external objection by providing a standpoint from which the rationality of commitments can be understood and assessed. Pro-

vided the commitment is rationally defensible, so too is the action that follows from it. Although rationality is itself always forward looking, it can provide reasons for taking account of the past.

The essays in this volume respond to and develop some of these themes in David Gauthier's thought. Robert Brandom's essay "What Do Expressions of Preference Express?" casts doubt on the concept of preference that is shared by the traditional theory of rational choice and Gauthier's important modifications of it. As we have seen, the traditional theory takes that concept as primitive and as having its authority primitively, while Gauthier's account limits its claim to authority to those cases in which the behavioural dimensions of an agent's preferences match the attitudinal ones. Drawing on Gauthier's distinction between these two dimensions of preference, Brandom opens up the question of what is expressed when someone expresses a preference. He argues that expressions of preference differ from the dispositions to choose that are behavioural preferences, because expressions of preference have propositional content. Brandom makes two claims about this content. He argues first that propositional content is necessary if preference sets are to be assessed for their rationality (or irrationality), because unless they have such content, sets of preferences do not stand in the relations of incompatibility that are presupposed by any evaluation. Second, he argues that the content must be understood in terms of the idea of a commitment to choose, rather than a mere disposition to do so. Drawing out the implications of this idea leads him to defend what he calls "minimal Kantianism" about normativity – the idea that values or norms are reasons only insofar as they are acknowledged as such by agents. On this view, reasons are not reducible to facts about agents. Brandom concludes that the concept of preference that is of interest to practical philosophy presupposes the idea of a reason, rather than explaining that idea.

Arthur Ripstein's essay "Preference" examines parallels between the role of preference in much recent moral philosophy and the role of preference in classical empiricism. Empiricist epistemology and utilitarian and contractarian moral philosophy have a common origin, and, so Ripstein would have it, a common weakness. Each seeks to account for a problematic concept – physical objects and knowledge in one case, a person's good or what one has reason to do in another – in terms of what is taken to be an unproblematic concept – sensation in the former case, preference in the latter. Each thus retreats to what appears to be a subjective account of the concept in question. The difficulties with empiricist accounts of perception are by now widely acknowledged; Ripstein seeks to show how the same problems undermine preference-based accounts of practical reason and goodness. The basic strategy of the essay is to show how the uncontroversial role of one's tastes in evaluating and justifying one's choices presupposes an independent account

of what one has reason to do, in much the same way that the uncontroversial role of perception in evaluating and justifying beliefs presupposes an independent account of the reliability of the agent's perceptual apparatus, and so an account of what the world is like. In a new concluding section of the essay, he shows that the difficulties of empiricist accounts do not lend support to the rationalist view that is often thought to be the only alternative. Instead, the failures of empiricist accounts of practical reason reveal the sense in which normative concepts cannot be reduced to factual concepts of any sort.

In her contribution, "Rational Temptation," Claire Finkelstein explores certain tensions she finds in Gauthier's preference-based instrumentalism about practical rationality. She argues that he cannot free himself as easily as he wants from certain properties of the economist's or decision theorist's notion of preference. Accepting the economic understanding of choice as constrained by preference commits one, she argues, to regarding counter-preferential choice as irrational. Thus she contends that although Gauthier's account offers an explanation of the rationality of undertaking commitments, it must always regard acting on commitments, and so acting counter to one's preferences, as irrational. That is, although theorists like Gauthier seek to make room for plans and intentions in order to enable agents to better satisfy their preferences, they may be unable to do so given their acceptance of the received view that preferences constrain rational choice. In the end, Finkelstein urges a view that relaxes the constraints imposed by preference on practical deliberation.

In "Bombs and Coconuts, or Rational Irrationality," Derek Parfit develops some of his critical reactions to Gauthier's revisionist account of practical rationality. Exploring a series of examples of rationally motivated irrationality, Parfit tries to show that rather than supporting the claim that morality is rational, Gauthier's arguments show only that it is sometimes advantageous to believe that it is. While it may be in our interest to have a certain disposition to act against our interests, and while it may be rational to bring ourselves to have this disposition, acting on it will still be irrational.

John Broome asks whether intentions are reasons and argues that they are not. He endorses what has come to be known as "the bootstrapping objection," due to Michael Bratman: If something is not a reason, it does not become one simply because an agent takes it to be one (and if it is a reason, the fact that an agent has formed an intention to act on it does not provide a further reason to do so). Reasons cannot be created out of nothing, as it were. Broome addresses some intuitive objections to his account and at the same time defends his account of the normative relationship between intending and acting. Intentions are, he thinks, normative requirements of some sort. But they are unusual requirements, not least because it is often permissible to change one's mind. Then, Broome argues, the original intention must be

repudiated. He argues that if one intends to do something and one does not repudiate this intention, the intention normatively requires one to do what one intended. His concluding discussion shows how his account addresses the problem of reasoning about incommensurate alternatives.

Michael Thompson's long and probing essay, "Two Forms of Practical Generality," explores the role of generality in normative practice. He displays a common logical structure in three kinds of examples: the idea, familiar from discussions of rule-utilitarianism, that an act can be justified by showing it to be an instance of a more general practice that is itself justified in some other way; Gauthier's idea that it can be rational to dispose oneself to act in certain ways; and the idea, familiar from discussions of promising, that having made a promise, one must put aside considerations that ordinarily would be sufficient to justify acting differently. In particular cases, the demands of the practice may come into conflict with the demands of the considerations that serve to justify the practice. Thompson explores the logical structure common to rules, practices, and dispositions that enables them to justify particulars. He shows that although they are not themselves substantive principles of morality or rationality, the "transfer" principles that allow an act to inherit its justification from a practice apply only to practices that have a certain sort of generality that cannot be characterized either sociologically or psychologically. Instead, they have a specific logical structure that is itself an expression of a fundamental feature of practical reason.

Adam Morton wishes to put aside questions of "rationality" for a moment and inquire about the *psychology* that is needed by maximizing agents whose interests lie in cooperation. Psychologies, in the sense that is of concern to Morton, are learned early in life as we pick up the doctrines, habits, and cognitive tricks of our culture. There is reason to think that these are connected to the patterns of interaction and cooperation in a culture. Morton asks what it would mean for a psychology in this sense to fit well with cooperative practices. The results, he conjectures, not only are instructive for agents such as ourselves but also allow us to avoid some of the counsels of despair of abstract decision theory.

Contemporary game theory has revealed the striking complexity of strategic human interactions. In earlier work, Peter Danielson has argued that many moral constructivists like Gauthier oversimplify the strategic choices facing rational agents who are choosing principles or dispositions to guide their actions. In this work, he deploys evolutionary game theory and computer simulations ("evolutionary artificial morality") to test which of several competing principles might prove most efficient for agents seeking to maximize utility. Modelling the interactions of sophisticated rational agents – capable of constraining their actions in complex ways – turns out to be even more daunting when one makes the choice of a principle or disposition part of the

game itself. In his contribution "Which Games Should Constrained Maximizers Play?" Danielson considers a number of ways of representing the interactions of different kinds of cooperative or principles agents. He raises a number of considerations about different information conditions and suggests that it may be rational for agents to reveal less about themselves than many theorists have argued.

In "The Strategy of Cooperation," Edward McClennen argues that social theory has mischaracterized ideally rational and knowledgeable agents. He thinks that rather than choosing how to act strategically in all situations, such agents can come to view their interactions with others as defining a practice calling for principled or rule-governed choice. He takes seriously some mid-twentieth-century remarks by Thomas Schelling about reorienting game theory and argues that rational agents in many situations should seek to coordinate on mutually beneficial outcomes rather than sub-optimal or inefficient equilibria. Agents who are able to make rule-governed choices should do better in many contexts than should the straightforward maximizers of orthodox social theory.

This volume concludes with "We Were Never in Paradise," a critical essay by Candace Vogler, which examines the place of practical reason in moral life by considering whether or not selves are sufficiently unified for talk about rationality and commitment to apply to their lives. This theme, prominent in much "postmodern" thought, is developed by Vogler in ways that make it accessible to, and reveal its fundamental importance for, philosophers working in the analytic tradition of Anglophone philosophy. Through an engagement with Rousseau's writings on the self, she seeks to undermine confidence in the rationalist and liberal conception of the person that is prominent in contemporary philosophy. In its place she articulates a Rousseauian idea of a self lacking in unity, yet still subject to moral demands.

Notes

1. See Robert H. Frank, *Passions Within Reason: The Strategic Role of the Emotions* (New York: Norton, 1988), ch. 5–7.
2. See especially David Gauthier, "Assure and Threaten," *Ethics* 104 (4): 690–721, 1994.