By “deciding how to decide”, I mean using practical reasoning to regulate one’s principles of practical reasoning. David Gauthier has suggested that deciding how to decide is something that every rational agent does. Whether or not we agree with Gauthier about agents in general, we might think that his suggestion applies well enough to many of us moral philosophers. We assess rival principles of practical reasoning, which tell us how to choose among actions; and assessing how to choose among actions certainly sounds like deciding how to decide.

One of my goals in this essay is to argue, in opposition to Gauthier, that assessing rival principles of practical reasoning is a job for theoretical rather than practical reasoning. How to decide is something that we discover rather than decide.

The idea that our principles of practical reasoning can be regulated by practical reasoning is essential to Gauthier’s defense of his own, somewhat unorthodox conception of those principles. And although I do not endorse the specifics of Gauthier’s conception, I do endorse its spirit. There is a flaw in the orthodox conception of practical reasoning, and Gauthier has put his finger on it. Unfortunately, Gauthier’s account of why it is a flaw, and how it should be fixed, ultimately rests on practical considerations, whose relevance is open to question if, as I believe, practical reasoning cannot regulate itself.

This essay therefore has a second goal, which complicates matters considerably. Although I want to reject Gauthier’s notion that we decide how to decide, I also want to preserve what rests upon that notion, in Gauthier’s view: I want to resettle Gauthier’s critique of the orthodoxy on a new foundation. I shall try to carry out this delicate operation as follows. First I’ll summarize Gauthier’s critique of the orthodoxy about practical reasoning. Then I’ll introduce Gauthier’s alternative conception of practical reasoning and his practical argument for deciding upon it. After explaining why I think that practical reasoning cannot be self-regulated in this manner, I’ll explain how I think that it must be regulated instead. Finally, I’ll return to Gauthier’s critique of the orthodox conception in order to reformulate it in theoretical terms.

GAUTHIER’S CRITIQUE OF STRAIGHTFORWARD MAXIMIZATION

The target of Gauthier’s critique is what he calls the theory of straightforward maximization. The theory of straightforward maximization says that an agent should choose, from among the discrete actions currently available to him, the one that yields the greatest expectation of benefit for him. Gauthier argues that the theory of straightforward maximization must be modified so as to enable rational agents to avoid falling into prisoner’s dilemmas.

Prisoner’s dilemmas get their name from a philosophical fiction in which two people—say, you and I—
are arrested on suspicion of having committed a crime together. The police separate us for interrogation and offer us similar plea bargains: if either gives evidence against the other, his sentence (whatever it otherwise would have been) will be shortened by one year, and the other’s sentence will be lengthened by two. In light of the expected benefits, maximizing rationality instructs each of us to give evidence against the other. The unfortunate result is that each sees his sentence shortened by one year in payment for his own testimony, but lengthened by two because of the other’s testimony; and so we both spend one more year in jail than we would have if both had kept silent.

The moral of this story might appear to be, not just that crime doesn’t pay, but that the pursuit of self-interest doesn’t pay, either. But of course the individual pursuit of self-interest does pay in this story, since each of us does better by testifying than he would have by keeping silent, irrespective of what the other does. What fails to pay, in this story, is self-interested action on the part of two agents, when compared with self-sacrifice by both. Joint sacrifice would have yielded greater benefits for each of us than joint selfishness.

In order to illustrate Gauthier’s complaint against the maximizing conception of practical reasoning, we must imagine that you and I have an opportunity to confer before being separated for interrogation. Since we expect to be offered incentives to betray one another, we try to attain solidarity by means of an agreement. “I’m willing to keep silent if you are,” I say, and you say, “I am, too.”[4] We thus appear to have agreed on joint sacrifice, to our mutual advantage.

Yet once we are led into our separate interrogation rooms and offered our separate plea bargains, the expected benefits of the alternatives are unaffected by our attempt at collaboration.[5] Each still stands to gain by testifying against the other, irrespective of what the other does; and so the principle of maximization still instructs each of us to testify, in violation of our supposed agreement.

What’s more, each of us could have predicted that the other would violate the agreement if only he had known that the other was a maximizer. And neither of us would have been willing to forgo his plea bargain in order to reach an agreement that the other was in any case going to violate. When I said “I’m willing to keep silent if you are,” I meant to express a willingness that would take effect in my behavior, but only on the condition that you express a willingness that would be equally effective in yours. Had I known that any willingness you might express was likely to be overridden by maximizing calculations, I would have realized that your expressed willingness to keep silent would be of no value to me, and so I would never have offered mine. Hence if either party’s allegiance to maximization had been known, no agreement would ever have been offered to him. Straightforward maximizers thus find themselves excluded from cooperative agreements.

Being excluded from cooperative agreements is a cost that counts against the maximizing conception of practical reasoning, in Gauthier’s eyes. Who, he asks, would want to have deliberative principles that would exclude him from the cooperative agreements by which prisoner’s dilemmas are avoided? In Gauthier’s view, one is better off adopting deliberative principles that favor the fulfilment of mutually beneficial agreements, so that one will be offered an opportunity to enter them.

Gauthier therefore proposes an alternative to the maximizing conception of practical reasoning. The alternative conception instructs an agent to maximize except when he is deciding whether to fulfil a commitment, in which case it compares the benefits, not of fulfilling or violating the commitment, but rather of two overall courses of action, one being that of making and fulfilling the commitment, the other being that of never having made the commitment at all.[6] The agent is instructed to fulfil his commitment if the package deal of making and fulfilling the commitment promises greater benefits than the package in which he never made it and so never had to consider fulfilling it.

But why should an agent base his decision on a comparison between these package deals? After all, an agent evaluates whether to fulfil a commitment at a juncture where never having made the
commitment is no longer an option: the commitment has already been made. Why should he care whether, in fulfilling the commitment, he will complete a course of action whose consequences are better than those of an alternative that is no longer available? And why should his evaluation at this juncture ignore a course of action that still is available—namely, that of making the commitment and then violating it? [7]

Gauthier’s answer to these questions is that the conception of practical reasoning that applies this comparison is superior to the straightforwardly maximizing conception, which compares all and only the discrete steps that are currently available. [8] What makes the alternative conception superior, Gauthier says, is that it enables the agent to make commitments that he can be counted on to fulfill, thus making him eligible as a partner in cooperative agreements. The benefits to be expected from cooperation give the agent reason for abandoning maximization and adopting Gauthier’s alternative conception instead.

Here is where the notion of deciding how to decide first enters Gauthier’s argument. His conception of practical reasoning is commended to the agent by practical considerations about the benefits of holding it. The deliberative principles that Gauthier favors are thus arrived at by deliberation.

The deliberations by which an agent arrives at these principles are articulated most fully in Gauthier’s paper “Assure and Threaten”. [9] There Gauthier explains that these deliberations are framed by an ultimate or overall goal of the agent’s—say, the goal of having as good a life as possible. In furtherance of this master-goal, Gauthier argues, straightforward maximization recommends its own replacement by the alternative conception, which will afford the agent a better life by affording him access to the benefits of cooperation.

At this point, however, we might suspect that an earlier problem has reemerged. Suppose that an agent adopts Gauthier’s conception of practical reasoning and is consequently offered a cooperative agreement, which he accepts. Holding Gauthier’s conception will potentially have furthered the agent’s master-goal by making the agreement available to him, since nobody would have offered to cooperate with a straightforward maximizer. But when the time for fulfilling the agreement arrives, Gauthier’s conception of practical reasoning will no longer further the agent’s goal, since his goal would be better served by reasoning that permitted him to violate the agreement. [10] If the agent’s conception of practical reasoning were an object of choice in the sense that it was continually open to revision, then the agent would now find himself changing conceptions in midstream.

Of course, if the agent had been expected to change conceptions in midstream, then the cooperative agreement would never have been offered to him, in the first place, since his conception of practical reasoning, being thus revisable, would have offered no guarantee of his future cooperation. Hence Gauthier’s conception of practical reasoning can create beneficial opportunities only if the agent appears unlikely to abandon that conception at a later date.

Gauthier has never explicitly addressed this problem, to my knowledge, but a solution to it is implicit in his discussions of the choice between conceptions of practical reasoning. The solution rests on the claim that Gauthier’s conception, unlike straightforward maximization, is self-supporting. [11]

We have envisioned that an agent, having adopted Gauthier’s conception of practical reasoning, might subsequently abandon it, at least temporarily, when he no longer benefits from holding it. But to abandon Gauthier’s conception of practical reasoning on the grounds that it is no longer beneficial would be to apply the principle of straightforward maximization; whereas the agent, having adopted Gauthier’s conception, is no longer a straightforward maximizer.

The question before the agent, at this juncture, is whether to retain a conception of practical reasoning,
and this conception can be regarded as a commitment to deliberate according to particular principles on practical questions. The question before the agent therefore belongs to the very class of questions on which his current conception of practical reasoning differs from straightforward maximization—questions, that is, about whether to abide by a previously adopted commitment. The agent’s current conception of practical reasoning recommends that he answer such a question in accordance with a comparison between the consequences of making and abiding by his commitment, on the one hand, and the consequences of never having made it, on the other, each considered as a package deal. Reasoning in this manner, the agent will find that the advantages of being offered a cooperative agreement outweigh the disadvantages of fulfilling it; and so he will decide to abide by his commitment, and hence to retain Gauthier’s conception of practical reasoning.

Thus, Gauthier’s conception is stable in a way that straightforward maximization is not. If a straightforward maximizer applies his deliberative principle to a comparison between itself and Gauthier’s principle, it will lead him to adopt Gauthier’s instead; but if he then applies his new deliberative principle to the same comparison again, it will lead him to stick with Gauthier’s. Indeed, the benefits that attracted him to Gauthier’s principle will have depended on the fact that it wouldn’t lead to its own abandonment, since its stability in this respect is what made him eligible as a partner in cooperative agreements.

The notion of deciding how to decide has now entered Gauthier’s argument at two distinct points. A straightforward maximizer will replace his conception of practical reasoning, according to the argument, because an alternative conception is more conducive to his master-goal and is thus recommended by practical reasoning as he now conceives it. And the alternative conception of practical reasoning will better serve the agent’s master-goal only because, once adopted, it will be retained at the recommendation of practical reasoning as he then conceives it.

**GAUTHIER’S ARGUMENT FOR DECIDING HOW TO DECIDE**

At both points the argument depends on the assumption that how to conceive of practical reasoning is a question to be settled by practical reason. Yet evaluating conceptions of practical reasoning need not be a practical matter.

What if there is an objectively correct way to deliberate—one principle, or set of principles, whose application is valid independently of its pragmatic advantages or disadvantages. In that case, the way to deliberate will be the correct way to deliberate, and the correct way to deliberate will not depend on how we would choose to deliberate if choosing on pragmatic grounds. It will rather be a theoretical matter, which depends on which conception is right.

Now, if evaluating conceptions of practical reasoning is not a practical but a theoretical inquiry, which seeks principles that are right independently of what we would prefer, then it won’t be an inquiry of the sort to which a conception of practical reasoning would apply. And if it is not an inquiry to which such a conception would apply, then no conception of practical reasoning will militate either for or against itself as the conclusion of that inquiry. A conception of practical reasoning cannot lay down principles that would lead to its own adoption or rejection if its adoption or rejection doesn’t depend on practical reasoning.

In short, conceptions of practical reasoning cannot be self-supporting or self-defeating unless evaluating such conceptions is an instance of practical reasoning. And why should we think of it that way?

Gauthier answers this question by contrasting two pictures of how conceptions of practical reasoning should be evaluated. He first considers a picture in which actions are subject to an independent criterion of success, and deliberative principles can therefore be evaluated by their tendency to yield actions that meet this criterion. As Gauthier points out, this picture portrays the practical sphere as
analogous to the theoretical, in which beliefs are subject to the independent standard of truth, and principles of theoretical reasoning can be evaluated by their tendency to yield true beliefs.

What Gauthier might have added is that this picture portrays the evaluation of deliberative principles as a theoretical matter, since the objective criterion of success for actions entails a criterion of correctness for deliberation. In this picture, the correct way to deliberate is whichever way best tracks the criterion of success for actions, just as the correct way to theorize is whichever way best tracks the truth. The criterion of success for actions isn’t up to us, in this picture; the tendency of deliberative principles to track that criterion isn’t up to us; and so the way to deliberate isn’t up to us, either. Hence evaluating conceptions of practical reasoning isn’t an instance of practical reasoning.

Gauthier considers a specific version of this picture, in which the independent criterion of success for actions is conduciveness to the agent’s master-goal. According to this criterion of success, of course, Gauthier’s deliberative principle takes second place to straightforward maximization, since it often recommends goal-defeating actions, such as the fulfilment of cooperative agreements. Not unsurprisingly, then, Gauthier prefers a different picture of how the agent’s master-goal bears on his conception of practical reasoning.

~

In Gauthier’s picture, the agent’s master-goal sets a standard of success for deliberative principles, not for actions, and it bears on actions only indirectly: the rationality of actions depends on their issuing from the principles that best promote the goal. Deliberative principles can promote the master-goal not only through the actions that they recommend but also through collateral effects such as creating opportunities for action—for example, by making the agent eligible as a partner in cooperative agreements. On this standard, Gauthier’s conception of practical reasoning surpasses straightforward maximization.

This picture favors Gauthier’s conception of practical reasoning precisely because it portrays the evaluation of such conceptions as a practical matter. In this picture, there is nothing about actions that deliberative principles attempt to track in their recommendations, and so there is nothing to make them objectively correct or incorrect. Conceptions of practical reasoning must therefore be evaluated, not by whether they get the principles of deliberation right, but rather by the pragmatic pros and cons of holding them.

What reason does Gauthier offer for favoring the latter picture? He says that he favors it because it doesn’t lead us to choose a conception of rationality that’s self-defeating, as straightforward maximization appears to be. “[I]t is surely mistaken,” he says, “to treat rational deliberation as self-defeating, if a non-self-defeating account is available.”[14]

~

But now Gauthier has argued in a circle. He is currently comparing, not just rival conceptions of practical reasoning, but rival pictures of how conceptions of practical reasoning should be evaluated. In one picture, conceptions of practical reasoning are evaluated by whether their deliberative principles track actions that meet the relevant criterion of success, which (Gauthier envisions) is conduciveness to the agent’s master-goal; and this evaluation militates against Gauthier’s conception. In the other picture, conceptions of practical reasoning are evaluated by whether holding them is conducive to the agent’s master-goal, and this evaluation militates in favor of Gauthier.

The question is which picture to adopt. Gauthier’s answer is that we should adopt the latter picture, since the former recommends a self-defeating conception of practical reasoning. But as we have seen, a conception of practical reasoning cannot be self-defeating or self-supporting unless it is self-applicable—that is, unless its own evaluation is an inquiry of the sort to which the conception itself applies. And in the picture that Gauthier rejects, evaluating conceptions of practical reasoning is not a practical but a theoretical inquiry, which seeks to ascertain which principles of deliberation are correct.
How can Gauthier reject this picture on the grounds that it favors a self-defeating conception of practical reasoning? The picture itself portrays conceptions of practical reasoning as incapable of defeating themselves, because they do not apply to their own evaluation. Only in the picture adopted by Gauthier do conceptions of practical reasoning apply to their own evaluation and thereby qualify as self-supporting or -defeating.

We have to join Gauthier in adopting this picture, then, before we can see it as saving us from a self-defeating conception of practical reasoning. Gauthier’s reasons for adopting the picture are therefore visible only after we have already adopted it.

What’s worse, Gauthier’s picture severely limits the normative force that’s available to conceptions of practical reasoning. Of course, any conception of practical reasoning will be normative in content, simply by virtue of applying terms like ‘reason’ and ‘rational’: any conception of practical reasoning will tell us how to deliberate and, by extension, what to do. But precisely because any conception will tell us how to deliberate, and hence what to do, we need to find a conception whose injunctions are authoritative or valid or genuinely binding. One problem is that Gauthier’s picture seems to rule out the possibility of our recognizing a conception as authoritative in this sense. If how to conceive of practical reasoning is itself a practical question, as Gauthier claims, then we shall have no conception of how to answer it until we have a conception of practical reasoning; and so we shall have no conception of how to answer the question until we have already answered it. We shall therefore find ourselves either unable to answer the question at all or forced to answer it arbitrarily. Having chosen a conception arbitrarily, we shall be equipped to reconsider our choice, of course, but only in an arbitrarily chosen manner. The results of such a procedure are unlikely to inspire confidence.[15]

This methodological problem may be tolerable by itself. Inquiry has to start somewhere, usually with received opinions, and it often relies on these opinions even in the process of criticizing and revising them. The principle of straightforward maximization may simply be the received opinion with which inquiry into practical reasoning is obliged to begin.

Yet Gauthier’s picture portrays more than a methodology for answering the question how to conceive of practical reasoning; it portrays the very nature of that question, as one to which no one answer is better than another except in so far as it is favored by practical reasoning. The picture consequently undermines, not just the possibility of our recognizing a conception as authoritative, but the very possibility of a conception’s being authoritative, in the first place. The only authority available to a conception of practical reasoning, in Gauthier’s picture, lies in the fact that the conception is supported by itself or by another conception that cannot boast even that much authority.[16] Why should we feel bound by principles whose only claim on us is that they are recommended by themselves or by other principles that have even less to recommend them?

For these reasons, I am inclined to prefer the picture that Gauthier rejects, in which conceptions of practical reasoning are evaluable in relation to an independent criterion of success for actions. The philosophy of rational choice must therefore begin by finding the criterion of success for actions, in relation to which conceptions of rationality can be evaluated.

One might think that this task will draw us into reasoning just as circular as that involved in Gauthier’s picture. A criterion of success for actions would seem to embody a normative judgment about how one ought to act. Finding a criterion of success for actions would therefore seem to be a practical inquiry about how to act, an inquiry that requires us to deliberate and hence to have deliberative principles already in hand. How, then, can we expect our criterion of success for actions to guide us in choosing deliberative principles?

A THEORETICAL FOUNDATION FOR DELIBERATIVE PRINCIPLES
The answer to this question is provided, I think, by the analogy between theoretical and practical reasoning—the analogy that Gauthier declines to apply. In the case of theoretical reasoning, our criterion of success for beliefs doesn't embody a normative judgment on our part; rather, the criterion of success for beliefs is determined by the internal goal of beliefs themselves.

As Gauthier observes, it is in the very nature of beliefs to aim at being true. Propositional attitudes that do not aim at the truth simply don't constitute beliefs, whereas attitudes that constitute beliefs do so partly in virtue of having that aim. Hence being true is simply what would be required for beliefs to succeed in their own terms. That beliefs must be true in order to succeed is a fact about them, given their goal-directed nature; it's not a normative judgment at which we arrive by practical reasoning. What is to count as success for belief is not for us to decide, because it's determined by an aim that's internal to belief itself.

Similarly, we can avoid circularity in our account of practical reasoning by finding an internal aim in relation to which actions can be seen as succeeding or failing in their own terms. If actions have a constitutive aim, then they will be subject to a criterion of success that's determined by their nature rather than our practical reasoning; and so that criterion will be available in advance of deliberation, as a basis for evaluating deliberative principles.

Some deliberative principles will then bear objective authority as norms of practical reasoning. There will be an objective fact as to what makes action successful as action, just as there is a fact as to what makes belief successful as belief. And in so far as deliberative principles tend to issue in action that succeeds as action, they will qualify as objectively correct ways to regulate action; just as theoretical principles qualify as correct ways to regulate belief in so far as they tend to issue in true belief, which succeeds as belief. The normative authority of deliberative principles will thus rest in the nature of action as constitutively directed at a particular aim.

Now, there are two possible views on the relation of aims to actions. One view is that there is no single aim that's constitutive of action in general, as aiming to be true is constitutive of belief. Actions are utterly heterogeneous as to their goals, according to this view: what's constitutive of action is simply having some goal or other.

This view yields a thoroughgoing instrumentalism about practical reason—an instrumentalism far more thorough, in fact, than that expressed in the norm of maximizing utility or value. For if the only goal internal to an action is the peculiar goal at which it is expressly directed by its agent, then each action will have to be judged a success or failure solely on the basis of its conduciveness to its peculiar goal. And very few actions are expressly directed by their agents at maximizing some overall measure of value. Most actions are directed at less ambitious aims, which would be the only aims in relation to which they could be said to succeed or fail in their own terms. A conception of practical reasoning would therefore have to be evaluated by its tendency to recommend actions that succeeded in promoting whatever they were severally intended to promote.

The alternative view is that there is a common goal that's constitutive of action—something at which behavior must aim in order to qualify as an action, just as a propositional attitude must aim at being true in order to qualify as a belief. According to a tradition that stretches from Aristotle to Davidson, action has a constitutive aim of this sort—namely, the good. Whatever an action aims at, according to this tradition, it aims at sub specie boni: in the guise of a good. In aiming at different things, actions are still aiming at things under the same description—as good—and so they share a constitutive aim.

This tradition might seem to offer a foundation for the maximizing conception of practical reasoning. If every action aims at something conceived as good, then conduciveness to the good would appear to be an internal criterion of success for actions—the criterion that actions must meet in order to succeed.
in their own terms. Since the most successful actions, by this criterion, are the ones that are most conducive to the good, the way to reach the truth about the success of actions will be to apply the method of maximization.

I believe that there are several errors in this foundational argument for the maximizing conception of practical reasoning. To begin with, aiming at things conceived individually as goods does not necessarily entail aiming to maximize a unified measure of goodness, since it doesn’t necessarily entail regarding one’s various ends as good in commensurable ways. More importantly, however, I doubt whether actions necessarily aim at things in the guise of goods.[20]

Yet the thesis that actions constitutively aim at the good is not the most plausible implementation of the view that they share a constitutive aim. This thesis assigns a constitutive aim to actions by claiming that they converge in their aims, in the sense that their various ends-in-view are all sought as means to, or components of, a single ulterior end. But the claim that actions converge in their aims is not necessary to the view that they share a constitutive aim.

The relation between the shared, constitutive aim of actions and their differing ends-in-view need not be that they aim to attain the former by attaining the latter; it may instead be that they aim to attain the former in the course of pursuing the latter. The constitutive aim of action, in other words, may be a goal with respect to the manner in which other goals are pursued, rather than a composite or expected consequence of those other goals.

Consider, for example, the goal of efficiency. You cannot pursue efficiency by itself, in a vacuum: you must have other goals in the pursuit of which you seek to be efficient. But then you can seek to be efficient at everything you do in the pursuit of those other goals. And in that case, efficiency is a goal of all your actions, not because you hope to attain it by attaining the various goals peculiar to those actions, but rather because you hope to attain it in the course of pursuing them.

I’m not claiming that efficiency is the constitutive goal of action: Lord, no. I’m just pointing out that actions can share a common aim even though their ends-in-view do not in any sense converge. Your actions can be aimed in as many different directions as you like and yet share the common aim of efficiency. My view is that there is a goal that is similar to efficiency in just this structural respect and that is shared by all your actions as such.

Identifying the relevant aim is not on my agenda for the present chapter.[21] My reason for introducing the possibility of such an aim is simply to fill out the methodological picture that I am opposing to Gauthier’s. In Gauthier’s picture, how to decide between actions is a practical matter, which we decide by weighing the practical advantages and disadvantages of various deliberative principles. In my picture, the principles for deciding between actions are not for us to decide, because they are determined for us by the point of action itself.

Gauthier’s picture suggests that we can hope for a pure theory of practical reasoning. Nothing underwrites our principles of practical reasoning, in his picture, other than the principles themselves. But my picture suggests that the philosophy of practical reasoning cannot be purified, in particular, of considerations from the philosophy of action. The correct principles of practical reasoning are determined by the constitutive aim of action—the thing at which behavior must aim in order to qualify as action, in the way that propositional attitudes must aim at the truth in order to qualify as belief. What makes an action rational, then, must depend on what makes something an action, to begin with. The philosophy of action must provide a foundation for the philosophy of practical reasoning.

Of course, the order of logical dependence need not dictate the order of discovery. Rival conceptions of practical reasoning can be taken as proposals for what the underlying aim of action might be. In any
deliberative principle, we can look for the implicit criterion of success that it tends to track, and we can ask whether satisfaction of that criterion might be the aim in virtue of which behavior qualifies as action.

Indeed, I think that Gauthier's critique of straightforward maximization can be reformulated along these very lines. What the maximizer's approach to prisoner's dilemmas reveals is, not that his conception of practical reasoning is disadvantageous, but rather that the criterion of success that it tracks in actions—namely, maximizing value for the agent—doesn't express a shared aim in which their status as actions could possibly consist.

**GAUTHIER'S CRITIQUE REFORMULATED**

In order to reformulate Gauthier's critique of maximization along these lines, I must return to our initial story, in which you and I expect to be offered incentives to testify against one another about a jointly committed crime. Gauthier's complaint against maximization is that it prevents us from reaching a mutually beneficial agreement to keep silent, since each of us would expect the other, as a maximizer, to violate such an agreement. Let me begin by re-examining this complaint.

The costs of being a maximizer are, in fact, less than obvious. Some philosophers have pointed out, for example, that the costs identified by Gauthier are primarily due, not to one's being a maximizer, but to one's being perceived as a maximizer; and they have argued against Gauthier's assumption that others can tell what one's conception of practical reasoning is.[22] But I want to grant Gauthier's assumption of translucency, as he calls it.[23] What is of greater interest for my purposes is that even translucent maximizers will not be excluded from cooperative agreements so long as they are capable of making commitments that are truly effective.

Suppose that when each of us says "I'm willing to keep silent if you are," what he thereby makes translucent to the other agent is a disposition that will withstand any subsequent calculations—an irrevocable disposition to keep silent under the specified conditions. In that case, forming and expressing a willingness to keep silent will be tantamount to making an effective commitment, which one's practical reasoning cannot overturn.[24] And the maximizing conception of practical reasoning will direct us to exchange such commitments, each conditional upon the other, since a conditional commitment will irrevocably dispose the issuer to keep silent only if the other agent has issued a reciprocal commitment, whose condition it will satisfy, so that the other will be irrevocably disposed to keep silent as well. Since each commitment will irrevocably lead to silence upon being reciprocated, silence will reign, and the harms of mutual betrayal will be avoided.

Rational maximizers seem to be capable of reaching cooperative agreements, then, provided that they can commit themselves effectively. So where are the costs of being a maximizer? And where is the need for a conception of practical reasoning that favors fulfilling cooperative agreements?

How, in fact, can there ever be a need for such a conception? The only commitment that either of us has an interest in making is one that is not only conditional on receipt of an effective commitment from the other but also sufficient to satisfy the condition on that reciprocal commitment, so that the other's silence will be ensured. And since each commitment will be conditional upon the efficacy of the other, each will have to be effective in order to satisfy the other's condition. Hence neither of us has any interest in making a commitment unless it is effective.[25]

So why do we need a conception of practical reasoning that will direct us to fulfit our commitments? The only commitments that we have any interest in making are effective commitments, sufficient to ensure their own fulfilment. A conception of practical reasoning that enjoined us to fulfil these commitments would thereby enjoin what is already guaranteed by the very existence of the commitments themselves. What would be the point?[26]
One might respond, on Gauthier’s behalf, that commitments of such efficacy are beyond the capacity of human agents. But surely human agents are capable of devising mechanisms for predetermining their future behavior; and in so far as they aren’t, this lack of ingenuity cannot be held against the maximizing conception of practical reasoning. A more plausible objection is that human agents do not ordinarily commit themselves by the brute predetermination of their behavior. The exchange of such mechanistic commitments wouldn’t add up to a cooperative agreement in any ordinary sense of the phrase.

Each of these commitments would be effective, we said, in the sense that it would govern the agent’s behavior irrespective of his subsequent deliberations about whether to fulfill it. Its efficacy would therefore consist in the fact that it left the agent no choice on that score. Yet if the agent would have no choice whether to fulfill his commitment, then his fulfilling it wouldn’t qualify as an autonomous action on his part. And cooperative agreements are ordinarily understood as leading, not just to coordinated behavior, but to cooperative action—that is, behavior whose coordination is wittingly and willingly sustained by the agents involved. The envisioned commitments wouldn’t enable rational maximizers to arrive at anything that would deserve to be called cooperative agreements in this sense.

Note that Gauthier’s objection to maximizing rationality has now quietly been transformed from the practical to the theoretical mode. The objection is no longer that someone who holds the maximizing conception thereby sacrifices the benefits of cooperative agreements. As we have seen, a rational maximizer can obtain those benefits so long as he can determine his future behavior by means of an effective commitment. The objection is now that although an effective commitment can help to produce the benefits of a cooperative agreement, it can’t help to constitute an agreement that would yield genuine cooperative action.

What’s wrong with the maximizing conception, then, is that, in so far as it can account for cooperation, it gets the nature of cooperative action wrong. And getting the nature of cooperative action wrong is not a practical but a theoretical failing—in particular, a failing in the philosophy of action.

Although Gauthier doesn’t charge the maximizing conception with a theoretical flaw of this kind, it is the very flaw that his conception of practical reasoning is suited to remedy. The need that’s filled by a conception favoring the fulfillment of cooperative agreements is the need to explain how we can rationally exchange commitments that will leave us autonomous in carrying them out.[27]

The commitments envisioned thus far threaten the agent’s autonomy because their efficacy depends on their power to withstand the agent’s own deliberations about whether to fulfill them; and their efficacy depends on this power because the agent is assumed to be a maximizer, whose deliberations will direct him to violate his commitments. But suppose that the agent held Gauthier’s alternative conception of rationality, which posits reasons for fulfilling mutually beneficial agreements. Then he could make commitments whose efficacy would consist precisely in the fact that he would subsequently be determined to fulfill them by his appreciation of reasons for doing so. And the agent’s being determined to act by an appreciation of reasons wouldn’t undermine his autonomy, since being determined by reasons is just what constitutes autonomy, according to compatibilism.[28] Rationally effective commitments wouldn’t have to withstand the agent’s subsequent deliberations about whether to fulfill them, thus leaving him no choice; rather, they would take effect through the agent’s deliberations and consequently through his choice.

In sum, a theory like Gauthier’s isn’t needed to provide rational agents with the benefits of cooperation; it’s needed to explain how rational agents can cooperate freely. The problem that Gauthier’s theory solves, but the orthodox theory doesn’t, is not the practical problem of how to achieve cooperation but...
the theoretical problem of how cooperating agents can be autonomous. This problem, I now suggest, is just an instance of a larger problem in characterizing the autonomy that rational agents enjoy over their future actions. The larger problem has nothing essentially to do with prisoner’s dilemmas or cooperation or, for that matter, morality in general. It simply has to do with our nature as agents who exercise the power of choice over future actions.[29]

Our autonomy over future actions requires, on the one hand, that we have the power of making future-directed decisions that are effective, so that we can determine today what will get done by us tomorrow. On the other hand, our future-directed decisions must not simply cause future movements of our bodies. If they did, our later selves would lack autonomy of their own, since they would find their limbs being moved by the decisions of earlier selves, as if through remote volitional control. We must exercise agential control over our own future behavior, but in a way that doesn’t impair our own future agential control.

The only way to control our future behavior without losing future control, I believe, is by making decisions that our future selves will be determined to execute of their own volition; and the only way to determine our future selves to do something of their own volition is by giving them reason to do it.[30] Hence future-directed intentions or commitments must be capable of providing reasons to our future selves. Unless we can commit ourselves today in a way that will generate reasons for us to act tomorrow, we shall have to regard our day-older selves either as beyond the control of today’s decisions or as passive instruments of them.

Yet the maximizing conception of rationality does not guarantee, or even make probable, that decisions made today will provide reasons tomorrow. According to that conception, the reasons available to our day-older selves will be generated by our day-older interests and day-older circumstances, which may or may not militate in favor of carrying out decisions made today. The maximizing conception of rationality therefore fails to account for the diachronic autonomy that we exercise in our future-directed decisions.

One might wonder whether there really is a problem here. If an agent forms an intention to do something in the future, and if he doesn’t change his mind, then the intention will remain in place and eventually come into the hands of his future self. When the intention subsequently produces an action, the agent’s future self will be acting of his own volition, since the intention producing the action will now be his.

Yet whether an agent acts of his own volition, when governed by an intention remaining from the past, depends on the manner in which it remains and governs. If the intention is simply a lit fuse leading to action by some self-sustaining causal mechanism that’s insensitive or resistant to the agent’s ongoing deliberations, then it is not really a volition of his current self; it’s just a slow-acting volition from his past. In order for the volition to become his own, the agent must buy into—or, at least, not be shut out of—its governance over his behavior.

Unfortunately, the practical reasoning of a straightforward maximizer will direct him to discard an intention whenever he stands to increase his expected benefits by doing so. An intention that doesn’t continue to serve his interests can remain to govern his behavior only by avoiding or resisting interference from his practical reasoning; and such an intention won’t qualify as his own volition. The only intention whose fulfilment will be attributable to the maximizer himself, and not to some mechanism within him, is an intention whose fulfilment serves his interests.

A capacity for such fair-weather intentions simply isn’t enough, in my opinion. Of course, we could circumvent the practical drawbacks of this capacity by adopting an axiology that guaranteed compatibility between the interests of our present and future selves.[31] In that case, commitments undertaken rationally would always be seconded by the rational deliberations of later selves, whose
interests would necessarily harmonize with those which made the commitments rational to undertake. But this axiology wouldn’t help to explain our actual capacity for future-directed commitments; it would merely wish away the real conflicts of interest in which the true nature of that capacity is revealed.

There are cases in which we have an interest in committing ourselves to future courses of action whose relation to our future interests is unknown, or even known to be adverse.[32] And experience tells us that we can often make rationally effective commitments in these cases. We can form resolutions that a future self will find rationally binding, whether or not they are seconded by his interests at the time.

Thus, our future-directed autonomy is not just a capacity to choose now what our future selves will in any case find reason for choosing then; nor is it a capacity to bind them to something else against their better judgment. It’s a capacity to make choices that our future selves will buy into but wouldn’t otherwise have made. And the straightforwardly maximizing conception of rationality cannot accommodate rational efficacy of this sort.

To be sure, the maximizing conception of rationality doesn’t rule out the possibility of giving our future selves reason to act. An agent may find or put in place arrangements whereby the interests that he will have tomorrow are somehow altered by the commitments he makes today. He can take out bets on his own constancy, for example, or he can train himself to feel costly pangs of self-reproach whenever he violates a past commitment. Making a commitment will then give his future self reason to follow through.

But these devices for conveying reasons to one’s future selves would be external to rational agency, as the maximizing conception portrays it. One could be a fully-fledged and perfectly rational agent without having any of these devices in place, and hence without having autonomous control over one’s future behavior. What’s more, employing these devices would entail treating one’s future selves as one treats separate people, since it would entail influencing their behavior indirectly, by modifying their incentives. If I offer you a large enough reward for following my directions, or threaten a large enough penalty for disregarding them, I put myself in a position to give you directions that will take effect without overriding your autonomy, but I do not thereby put myself in a position to decide what you are going to do. The ability to influence you by manipulating your expected pay-offs does not give me agential control over your behavior. Yet the control I would enjoy over my own future behavior via the devices under consideration would be no different.

Finally, some of these devices—indeed, the ones best able to resist being characterized as external manipulation—would depend on thoughts or feelings that the maximizing conception itself must regard as baseless. Training oneself to feel bad about violating commitments might enable one to undertake commitments that give one’s future selves reasons to act; but it would entail training oneself to feel bad about something that one has no reason to feel bad about, according to the maximizing conception, since that conception treats violating commitments as a perfectly rational thing to do.

For all of these reasons, the devices at the disposal of the maximizing conception of rationality provide at most a simulation of diachronic autonomy. They enable an agent to induce his future selves to act, but not in a way that amounts to deciding on future actions.

I believe that in failing to accommodate this form of autonomy, the maximizing conception misrepresents what action is. The criterion of success that maximizing principles are designed to track doesn’t express a constitutive aim for action, since behavior oriented toward satisfying that criterion wouldn’t amount to action, as we know it.

If an action were the sort of thing whose success or failure could be judged solely by utility-maximizing
considerations, then it wouldn't be the sort of thing that we could decide on today in a way that would necessarily give us reason to perform it tomorrow, and so it wouldn't be behavior over which rational agents had diachronic autonomy. But action is that sort of thing—it is behavior over which rational agents have diachronic autonomy—and so it can’t be the sort of thing whose success or failure can be judged solely by utility-maximizing considerations.

My view is that the constitutive aim of action—the aim in virtue of which behavior becomes action, and against which it can be judged a success or failure as action—is autonomy itself.[33] How behavior can aim at autonomy, and why it thereby qualifies as action, are questions beyond the scope of this chapter.[34] Here I can merely point out that if autonomy is the constitutive goal of action, and hence the internal criterion of success for action, then reasons for acting will be considerations relevant to autonomy, rather than considerations relevant to utility or the good. And we can at least hope that reasons of this kind will be generated by future-directed decisions. An analysis of action as behavior aimed at autonomy may therefore explain how future-directed decisions can govern future behavior rationally; and so it may explain how the nature of action makes diachronic autonomy possible.

The notion of autonomy as the constitutive goal of action isn’t essential to my present argument, however. Maybe other proposals for the constitutive goal of action would explain how considerations relevant to that goal forge an autonomy-preserving link between past decisions and future actions. My present argument invokes autonomy—specifically, diachronic autonomy—simply as a feature of action that cannot be explained by the orthodox conception of practical reason, because it gives no rational weight, and hence no autonomy-preserving influence, to past decisions.

There are of course those whose normative intuitions oppose giving rational weight to past decisions: they think that abiding by a commitment for its own sake is foolish. Then there are those whose normative intuitions demand rational weight for past decisions: they think that abandoning a commitment is fickle. But I am joining neither of these camps, since I cannot see how to settle the issue on normative grounds. What settles the issue, in my mind, is not an intuition to the effect that we ought or ought not to give weight to past commitments. What settles the issue for me is that simply being rational agents enables us to exercise autonomy over our futures without impairing our future autonomy—something that we couldn’t do unless past commitments had rational weight.

Gauthier’s critique of straightforward maximization has now been removed from its foundation in practical reasoning and resettled on a new foundation in the philosophy of action. I believe that the philosophy of action can provide a foundation, not just for criticizing inadequate conceptions of practical reasoning, but also for constructing an adequate conception. What it will thereby provide, of course, is not a foundation for deciding how to decide. How to decide is something that we will discover, by discovering what it is to act.

APPENDIX: COOPERATIVE AGREEMENTS

Why do I formulate each commitment as making the agent’s action conditional on the other agent’s commitment rather than on the other agent’s action? Consider.

A condition that an agent places on his action won’t be enforceable unless he can test for its satisfaction before he has to act. Of course, Alphonse can refuse to act unless Gaston acts first; but then Gaston mustn’t do likewise, or action will never get started. And Gaston cannot refuse to act unless Alphonse acts afterwards, since Gaston can’t test for the satisfaction of this condition before he has to act. At best, Gaston can refuse to act unless something is present that will subsequently guarantee action from Alphonse. A transparently effective commitment on the part of Alphonse provides such an assurance.

Now, Alphonse shouldn’t just commit himself to act on the condition that Gaston acts first. For suppose that Gaston would act first anyway, whether or not Alphonse was committed to reciprocate. In that
case, Alphonse would unnecessarily have encumbered himself in advance with the costs of reciprocating, since he didn’t have to do so in order to gain the benefits of Gaston’s action. Alphonse should therefore commit himself to act on the condition that his being so committed is necessary to elicit action from Gaston.

But how can Alphonse test whether this condition is satisfied? If Gaston is to act first, what will make it the case that Alphonse’s commitment was necessary to elicit his action? And how will Alphonse tell whether it was the case? The only way for it to be the case that Alphonse’s commitment was necessary, and for Alphonse to tell that it was, will be for Gaston to have previously framed a transparent commitment making his action (bi)conditional on Alphonse’s commitment. If Alphonse wants to lay down an enforceable condition, then, he should commit himself to act on the condition that Gaston frame a commitment that’s conditional on his, Alphonse’s, being so committed.

Of course, if Gaston is to fulfil this condition by framing a commitment, he will want to avoid thereby encumbering himself with the costs of acting unless his doing so is necessary in order to elicit action from Alphonse. So he will want to commit himself to act on the condition that Alphonse’s commitment is conditional on his, Gaston’s, being so committed. The agents will thus issue mirror-image commitments, each conditional upon the other.

Yet neither agent needs to spell out his commitment at such length. Each wants his commitment to be conditional on its being required to satisfy the condition on a commitment framed by other. And this condition can be imposed by a commitment to act simply on the condition that the other agent frame a similar commitment, since a similar commitment from the other agent will be one whose condition this commitment is required to satisfy.

A further problem arises, however, as to whether this commitment is sufficiently determinate in content. When Alphonse says “I’ll act if you frame a similar commitment,” what he is undertaking depends on what would count as a similar commitment from Gaston; and yet what would count as a similar commitment depends on what exactly Alphonse is undertaking. Alphonse’s commitment is thus self-referential in a way that leaves its content ungrounded.[35]

Some philosophers have proposed a syntactic solution to this problem.[36] According to this solution, Alphonse’s commitment is conditional on Gaston’s framing a commitment that is syntactically similar—that is, framed in the same words. But this solution is clearly unsatisfactory, since it interprets Alphonse’s commitment as requiring him to cooperate so long as Gaston says “I’ll act if you undertake a similar commitment,” even if Gaston is speaking an idiolect in which ‘act’ means “refrain” or ‘similar’ means “different”.

I think that the solution to the problem of ungroundedness, in this instance, is to recognize that a similar commitment can be defined as a commitment with the same determinate or potentially indeterminate content. Suppose that Alphonse frames the conditional commitment to act if and only if Gaston issues a particular commitment; but that he risks failing to specify the requisite commitment because he attempts to specify it as similar in content to his own. And suppose that Gaston frames the conditional commitment to act if and only if Alphonse issues a particular commitment; but that he risks failing to specify the requisite commitment because he attempts to specify it as similar in content to his own. In that case, Alphonse and Gaston have issued commitments that really are similar, precisely because they share the same potential indeterminacies of content in addition to the determinate aspects of content that they share. The upshot is that the risk of indeterminacy in their commitments has been avoided, after all, since it is perfectly determinate whether either person’s commitment has the same potential indeterminacies as the other’s.

NOTES

1. This chapter originally appeared in Ethics and Practical Reason, edited by Garrett Cullity and Berys
Gaut (Oxford: Oxford University Press, 1997), 29–52. I received valuable comments on this chapter from Elizabeth Anderson, Jim Joyce, John Broome, and Stephen Darwall. It also had the benefit of an excellent commentary by Piers Rawling at the St. Andrews conference, as well as comments from participants, including Christine Korsgaard, Peter Railton, Joseph Raz, Michael Smith, Garrett Cullity, and Berys Gaut.


3. Note that this theory does not spell out specific procedures for applying this principle of choice; it simply states the principle that ought to be implemented in one’s deliberative procedures. All of the conceptions of rationality discussed below are articulated at the same level of generality, as principles of choice that will no doubt require more specific procedures for their implementation. To argue, as I do, that practical reasoning cannot regulate our principles of choice is not to argue that it cannot regulate the specific procedures in which those principles are implemented.

4. Note that I say, “I’m willing to keep silent if you are,” not “I’m willing to keep silent if you do.” For a discussion of how this commitment is formulated, see the Appendix to this chapter.

5. We can of course imagine mechanisms by which our agreement would have altered the expected benefits. For example, we might belong to a gang that exacts revenge on liars but not on stool-pigeons. But the argument depends on the assumption that no such mechanisms are in place.


12. Of course, there may be more than one specific procedure by which these principles can be applied. But multiple procedures, in this sense, will all count as one way of deliberating, according to the scheme of individuation that was adopted in n. 3, above.

13. The point made in this paragraph is made by Parfit on pp. 19–23 of Reasons and Persons.


15. Christine Korsgaard suggests that Kant’s purely formal conception of practical reason is a solution to this problem. Imagine that an agent is free to choose his conception of a reason for acting. The
current problem is that, in order to avoid adopting a conception arbitrarily, the agent would seem to need a reason for adopting one conception rather than another; and yet he has as yet no conception—or only an arbitrarily chosen conception—of what would count as a reason for doing so. But perhaps the agent can adopt, as his conception of a reason for acting, the mere form of a reason for acting. Or rather, he already has this much of a conception, insofar as he is already committed to adopting something that will qualify as a conception of a reason. Since nothing that failed to respect the form of a reason would qualify as a conception thereof, the agent who sets out to choose his conception of a reason already has this much of a conception, and he therefore needs no reason for choosing it. He can end up with this purely formal conception of a reason without any circularity or arbitrariness. So, at least, says Korsgaard’s version of Kant. ("Morality as Freedom", in Y. Yovel [ed.], Kant’s Practical Philosophy Reconsidered [Dordrecht: Kluwer, 1989], 23–48, at 30–1.)

16. This is the point that Parfit makes on pp. 19–20 of Reasons and Persons.


18. “Assure and Threaten”, 699. Actually, what Gauthier says is that “[t]o believe is to believe true.” But this remark fails to pick out a distinctive feature of belief, since to desire is to desire true, to intend is to intend true, and so on, for all propositional attitudes. I assume, however, that Gauthier means to pick out the distinctive relation that belief bears to the truth, which is that to believe is to believe true with the aim of thereby getting the truth right. On this point, see Chap. 4, above; Chap. 7, above; and Chap. 10, below.


20. This doubt is the theme of Chap. 4, above.

21. I defend the view that action has a constitutive aim in Chap. 5, above. I offer one account of this aim in Practical Reflection (Princeton: Princeton University Press, 1989); and Chap. 6, above. I offer a different (though, I believe, compatible) account in Chap. 7, above.


24. Here I seem to imply that the disposition to keep silent is first formed and then expressed. I am inclined to believe, instead, that the disposition in question is formed precisely by being expressed. (See Chap. 8, above.)

25. Here, as elsewhere, the argument relies on Gauthier’s assumption of translucency.

27. Note the similarity to Rousseau’s formulation of “the fundamental problem” to be solved by the social contract: “How to find a form of association which will defend the person and goods of each member with the collective force of all, and under which each individual, while uniting himself with the others, obeys no one but himself, and remains as free as before” (The Social Contract, trans. Maurice Cranston [Harmondsworth: Penguin, 1968], I. vi, p. 60).

28. This point is made by Holly Smith, “Deriving Morality from Rationality”, n. 13.


30. This claim does not commit me to sharing Gauthier’s view on the so-called toxin puzzle. In the toxin puzzle, an agent is offered a large reward merely for forming a future-directed intention to do something mildly harmful to himself, such as drinking a toxin. The puzzle was originally offered by Gregory Kavka as a reductio ad absurdum of Gauthier’s theory, which implies that after the agent has collected the reward for intending to drink the toxin, he should follow through on his intention, even though he has something to lose by doing so and nothing more to gain (see Kavka, “The Toxin Puzzle”, Analysis 43 [1983] 33–36). Gauthier has simply bitten the bullet in this case, arguing that the agent should indeed drink the toxin (“Assure and Threaten”, 707–709).

Yet I needn’t endorse this course of action in order to hold that future-directed intentions provide reasons for following through. For one thing, my view may not require that the reasons provided by future-directed intentions outweigh all countervailing reasons; perhaps they should be required only to carry some rational weight. On this version of the view, the agent in the toxin puzzle may form an intention to drink the toxin, and thereby give his future self some reason to drink it, without giving him sufficient reason.

Of course, an intention that doesn’t provide sufficient reason for following through will be inefficacious, according to my view. The question therefore arises whether the reward is being offered only for an efficacious intention, or whether an inefficacious intention would do. A reasonable answer might be that an inefficacious intention wouldn’t be much of an intention and, in fact, might not be an intention at all. It that case, an agent’s intending to drink the toxin would require giving his future self sufficient reason for drinking it. Even so, I needn’t conclude that a rational agent will find himself with sufficient reason for drinking the toxin; I can conclude instead that a rational agent will be unable to muster an intention to drink the toxin, precisely because he’ll be unable to give his future self sufficient reason for drinking it. (This treatment of the toxin puzzle corresponds to Dan Farrell’s in “Intention, Reason, and Action”, American Philosophical Quarterly 26 [1989] 283–95.)

31. This point was made in discussion at the St Andrews Conference by Joseph Raz and Michael Smith.

32. Potential prisoner’s dilemmas are not the only such cases. For additional examples, see Edward F. McClennen, Rationality and Dynamic Choice: Foundational Explorations (Cambridge: Cambridge University Press, 1990).

33. Thanks to Christine Korsgaard for daring me to confess this view, in the discussion at St Andrews.

34. See Chap. 7, above. But see also Chap. 2.

35. Versions of this problem—or a problem similar to it—are discussed by Holly Smith, “Deriving Morality from Rationality”, 240–42, esp. n. 18; Peter Danielson, “Closing the Compliance Dilemma: How it’s Rational to be Moral in a Lamarckian World”, in Vallentyne (ed.), Contractarianism and Rational Choice, 307–15; and Richmond Campbell, “Gauthier’s Theory of Morals by Agreement”, 250–51. Some of these authors formulate the relevant commitments as making each agent’s action conditional on the other’s action. For them, the problem is how Alphonse tells whether he should act, given that he has made his action conditional on Gaston’s, which is in turn conditional on his,
Alphonse's. This problem disappears if the agents make their actions conditional on one another's commitments. But the problem of each action's being conditional on the other action is then replaced by the problem of each commitment's being dependent for its content on the other's content.


http://dx.doi.org/10.3998/maize.13240734.0001.001