existing moral motivations and psychological ties to other people, and in the absence of coercion and manipulation. The skeptic can still deny that these arrangements are impartial without qualification, since their appearance of impartiality depends on the existence of rough equality among people, and so she can deny that the arrangements count as *moral* arrangements. Yet her position seems to need qualification. The structure of interaction can make cooperative dispositions rational.

Yet I believe that Gauthier's showing can hardly be an adequate justification of any set of moral principles. Nor does the contractarian conception of the skeptical challenge seem one that can or need be answered. We know there are conflicts between morality and individual expected utility maximization, and this by itself is not sufficient to defeat morality. Nevertheless, it is appropriate to seek an understanding of how a moral code could be justified. We need to reconsider what a justification of morality could consist in, and to reconceptualize the skeptical challenge to morality.

Skepticism should be understood to demand a showing that some moral code as such can be justified. The skeptic arguably would be in difficulty if complying with a moral code, or disposing oneself to comply, were rationally justified under every possible circumstance for every agent, but she is not in difficulty if complying, or disposing oneself to comply, is merely justified for certain people under certain circumstances. The explosive device in the brain makes the point. A putative justification cannot be successful if it is grounded in idiosyncracies or special contingencies. This is a reason why justification by the explosive device is uninteresting. It is also a reason why justification cannot succeed on the basis of restrictive assumptions about preferences, such as nontuism, or restrictive assumptions about psychological characteristics, such as transparency. Contractarianism is a practical person-centered account of justification, and I claim that accounts of this kind, if grounded in the standard utility-maximizing theory of practical reason, or any subjective instrumental relativistic theory, *cannot* answer skepticism. For accounts of this kind are ultimately grounded in people's psychological states, such as preferences, which vary idiosyncratically from person to person. And accounts of this kind evaluate people's choices or dispositions, not a moral code as such.

It is hard to see how a nonidiosyncratic justification of some moral code as such could be achieved. If it is impossible within a practical person-centered theory, as I believe, then we need a new model of moral justification. It is unclear what this model could be like, unless we return to an epistemic conception. Our difficulty in knowing where to go from here reflects the unsatisfactory state of our understanding of the nature and justification of morality.

14. Deriving morality from rationality

Holly Smith

Introduction

From its earliest beginnings, western philosophy has attempted to forge a strong link between rationality and morality. Contemporary social contractarian theories derive a good deal of their attractiveness from their claim to have achieved this goal. Such theories argue that the principles of justice, or the principles of morality, issue from a contract that rational individuals would agree to initially, and would comply with once implemented. These theories present moral norms as issuing from rational choice, and so claim to establish the desired connection between morality and rationality.

David Gauthier is the most recent advocate of this approach in his attempt to provide a contractarian justification for moral behavior. In defending his enterprise, Gauthier explicitly invokes the desirability of establishing what he calls "the deep connection" between reason and morality.[1] He asserts that "The main task of our moral theory [is] – the generation of moral constraints as rational . . . ,"[2] and develops this thought in the following passage:

> . . . [T]he language of morals is . . . surely that of reason. What theory of morals, we might . . . ask, can ever serve any useful purpose, unless it can show that all the duties it recommends are also truly endorsed by each individual's reason? . . . But are moral duties rationally grounded? This we shall seek to prove, showing that reason has a practical role related to but transcending individual interest, so that principles of action that prescribe duties overriding advantage may be rationally justified. We shall defend the traditional conception of morality as a

[1] David Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986). p. 4.
[2] Ibid., p. 7.

> rational constraint on the pursuit of individual interest. . . . Our enquiry will lead us to the rational basis for a morality.[3]

Thus, it is essential to the success of Gauthier's project in his own eyes that he can indeed establish the link between rationality and morality. And in taking this stance, he joins a venerable western tradition advocating the importance of this link.

I shall argue, however, that Gauthier fails to show that morality is based on rationality. To see how his argument falls short, let me very briefly describe his main line of thought. Gauthier begins with Hobbes's and Rawls' idea that society is a "cooperative endeavor for mutual advantage." Human beings, living in isolation from each other, can only expect to do poorly in their struggle to survive and flourish. Social cooperation would enable each person to fare better than he or she could do in isolation. Unfortunately, the more common types of potentially profitable cooperation are ones in which individuals' direct pursuit of self-interest paradoxically produces an outcome in which each person is worse off than he would have been if everyone had acted less selfishly. Consider a case in which you and I are two fishermen inhabiting adjoining properties along a dangerous coastline. Hidden sandbars often cause our boats to run aground and our catch to be lost. Each of us can expect two such accidents in the coming year, one on our own sandbar, and one on our neighbor's sandbar. If either of us erected a lighthouse, it would prevent any accidents on the adjacent sandbar. The cost to each of us of a single accident is $500, whereas the cost per year of erecting and maintaining a lighthouse is $600. In these circumstances, if each of us considers only our own welfare, neither will build a lighthouse, since the annual cost exceeds the benefit by $100. But each of us fares worse under this arrangement (where we each suffer an annual loss of $1,000 from accidents) than under an arrangement in which both of us build a lighthouse, for each lighthouse benefits *both* its builder and her neighbor (if both build, each suffers a yearly cost of only $600).

The dilemma posed by this situation can be represented in the following standard diagram:

		I	
		Build	Not build
You	Build	– 600 – 600	– 500 – 1,100
	Not build	– 1,100 – 500	– 1,000 – 1,000

In classic Prisoner's Dilemmas such as this one, an optimal outcome cannot be reached by each agent pursuing his or her own interest. It

could be reached, however, if behavior were constrained by principles prohibiting purely selfish behavior: for example, a principle requiring each fisherman to build a lighthouse. According to Gauthier, such principles, if they impartially constrain the pursuit of direct self-interest, qualify as *moral* principles. But Gauthier also argues that it is *rational* for each individual to adopt and comply with such principles – rational in the standard sense of maximizing one's own self-interest. The core of Gauthier's project is to argue for this thesis, which if true would establish morality as part of the theory of rational choice.

When mutually beneficial outcomes are available through cooperation, but purely selfish behavior would disadvantage everyone, individuals have reason to work together to secure mutual benefits. However, many different cooperative arrangements are often available in a given case. For example, one cooperative arrangement just described in the fishermen's case has each of us building her own lighthouse; but a second arrangement, also to the advantage of each, would have each of us building her own lighthouse, but you paying me $50 to defray my expenses; a third would involve my instead paying you $50 to defray your expenses; and so forth. Individuals who might benefit greatly under one arrangement would do far less well under an alternative arrangement that favored others. In these circumstances, members of the group must bargain with each other to determine which particular arrangement they will adopt. Gauthier argues that rational, fully informed people will bargain according to what he labels the "Principle of Maximin Relative Benefit." Bargains reached in accord with this principle mandate cooperative arrangements in which the benefits created by cooperation are distributed so that the return to each individual is proportionate to the contribution (under some interpretation) that he or she brings to the cooperative enterprise. Individuals will bargain with each other to establish the basic terms of their future social interaction – the principles of morality and justice that will govern them – and will agree to norms that require distributing the benefits of interaction proportionally to the contributions of each interactor. Thus, Gauthier argues that they would agree to norms requiring promise keeping, truth telling, and fair dealing, because adherence to such norms permits people to cooperate in ways that may be expected to render the benefits of concrete interactions proportionate to the contribution provided by each party.

People in the state of nature, bargaining with each other to establish mutually beneficial constraints on purely selfish behavior, must look to the future in arriving at their agreement, in the sense that they recognize it is pointless to agree on constraints with which no one will later comply. Thus, Gauthier must show that people would indeed comply with the constraints – that is, the moral principles – agreed to in the original bargain, for otherwise no bargain will be struck.

Hobbes was driven to solve this compliance problem by invoking an

[3] Ibid., pp. 1, 2.

all-powerful sovereign to enforce terms of the bargain by threats of harm to disobedient citizens. Many contract theorists have agreed with Hobbes that compliance cannot be secured without coercive measures of some kind. Perhaps Gauthier's most distinctive contribution to contractarian theory is his argument that coercion is unnecessary to secure compliance. He argues that it is rational for a would-be bargainer to "dispose himself" to comply with the terms of the initial bargain as long as he expects similar compliance from others. Agents so disposed may occasionally be exploited by others who take advantage of their willingness to comply with agreements made. But, Gauthier argues, given plausible assumptions about human abilities to ascertain each others' dispositions – assumptions that human beings are "translucent," although not completely "transparent," to each other – a person who disposes herself to comply with fair bargains will enjoy opportunities for beneficial cooperation that will be denied to persons lacking this disposition. Because of these opportunities, she maximizes her self-interest by disposing herself to comply. And, Gauthier argues, having disposed herself to comply, it is rational for her actually to comply when the occasion arises, even when she would maximize her utility by violating the agreement. Thus, compliance with socially adopted norms can be secured by largely voluntary means, and each would-be cooperator knows there is point to agreeing to the bargain, since each will have reason to comply with the bargain once made.

Gauthier's core argument for the rationality of compliance

I begin by examining Gauthier's core argument for the rationality of compliance. To understand Gauthier's official statement of it, we must first introduce his definitions of several key terms. First, Gauthier defines a "constrained maximizer" as

> (i) someone who is conditionally disposed to base her actions on a joint strategy or practice should the utility she expects were everyone so to base his action be no less than what she would expect were everyone to employ individual strategies, and approach what she would expect from the co-operative outcome determined by minimax relative concession; (ii) someone who actually acts on this conditional disposition should her expected utility be greater than what she would expect were everyone to employ individual strategies.[4]

In terms of our fisherman example, what this means is that you are a constrained maximizer if (i) you form the intention to build your lighthouse if I build mine, and the intention not to build your lighthouse if I do not build mine (since the utility you would receive if both you and I build our lighthouses exceeds what you would receive if neither of us built one); and moreover (ii) you actually *do* build your lighthouse if, and only if, you expect me to build mine (since your expected utility of building a lighthouse only exceeds what you would get if we both fail to build lighthouses in a case where you expect me to build mine as well). A *straightforward maximizer*, by contrast, is someone "who seeks to maximize his utility given the strategies of those with whom he interacts" – that is, a normal maximizer of expected utility.[5] In Gauthier's initial statement of his argument, he assumes that the agents involved are *transparent* to each other – that is, each is "directly aware . . . whether he is interacting with straightforward or constrained maximizers."[6] For the sake of greater plausibility, this assumption is then weakened to the assumption that the agents are merely *translucent* – that is, ones whose dispositions "to co-operate or not may be ascertained by others, not with certainty, but as more than mere guesswork."[7] Because it is simpler to state the argument using the stronger assumption of transparency, I shall initially stick with it.

Gauthier's own statement of his argument goes as follows:

> Suppose I adopt straightforward maximization. Then I must expect the others to employ maximizing individual strategies in interacting with me; so do I, and expect a utility, *u*.
>
> Suppose I adopt constrained maximization. Then if the others are conditionally disposed to constrained maximization, I may expect them to base their actions on a co-operative joint strategy in interacting with me; so do I, and expect a utility *u*'. If they are not so disposed, I employ a maximizing strategy and expect

[4] Ibid., p. 167. Gauthier uses the term "disposition" to describe constrained maximization. It is troubling that he never explains what this crucial term means. Since it is difficult to get a grip on the logic of the argument without an explication of "disposition," I have made the most natural assumption, namely, that a disposition is an intention to perform a certain kind of act. This explication preserves an important feature that Gauthier wants, namely, that the later choice to perform the act is genuinely a free act.

My informal formulation of constrained maximization (immediately following in the text) differs slightly from Gauthier's formal definition. Strictly speaking, Gauthier's (somewhat unclear) definition only stipulates what a constrained maximizer does if (in our case) she expects her partner to build her lighthouse; it does not stipulate what she does if she expects her partner *not* to build her lighthouse. However, it is clear from Gauthier's discussion that constrained maximization requires the person to form, and carry out, the intention not to build her lighthouse if she expects her partner not to build. It is this aspect of the disposition that protects the constrained maximizer from exploitation by straightforward maximizers.

My informal characterization of constrained maximization assumes the case of transparency, in which the agent knows that her partner will form, and carry out, the intention to build (or the intention not to build, as the case may be). In the case of translucency, in which the agent may assign probabilities less than 1 to each possible action of her partner, constrained maximization could be stated as "forming and carrying out the intention to build if the expected utility to her of doing so exceeds her expected utility if both partners fail to build; and forming and carrying out the intention not to build otherwise."

[5] Ibid., p. 167.

[6] Ibid., pp. 173–4.

[7] Ibid., p. 174.

u as before. If the probability that others are disposed to constrained maximization is p, then my overall expected utility is $[pu' + (1 - p)u]$.

Since u' is greater than u, $[pu' + (1 - p)u]$ is greater than u for any value of p other than 0 (and for $p = 0$, the two are equal). Therefore, to maximize my overall expectation of utility, I should adopt constrained maximization.[8]

Having argued that it will maximize an agent's expected utility to adopt constrained maximization, Gauthier then claims that *since* it is rational to choose to be a constrained maximizer, it is *also* rational to carry out that choice – that is, to cooperate when the time comes.[9]

Let us lay out this argument, somewhat less technically, in terms of our example of the fishermen. We may envision the situation as a symmetrical one, in which each of us is faced with the same choice. The argument may be stated as follows:

1. I can now choose between constrained maximization (CM) or straightforward maximization (SM), and my choice will bring about my actually carrying out the chosen option. The options are:

   CM. Forming the intention to build if you build, or not build if you do not; and then actually building if I expect you to build, or not building if I expect you not to build.

   SM. Forming the intention not to build whatever you do; and then actually not building whatever I expect you to do.
2. Each of us is transparent, so each will be aware of the other's choice, and will know that the other will actually carry out that choice.
3. If I choose CM, then
   (a) If you choose CM, you will build and I will build.
   (b) If instead you choose SM, you will not build, and I will not build either.
4. If I choose SM, then
   (a) If you choose CM, you will not build, and I will not build.
   (b) If instead you choose SM, you will not build, and I will not build either.

[8] Ibid., p. 172. This argument contains what appears to be a (recurrent) misstatement, namely, that others are "conditionally disposed" to constrained maximization. There is never any question of people being *conditionally disposed* to constrained maximization; constrained maximization is already itself a disposition to form certain conditional intentions, and to carry one of them out when the specified conditions occur. "A constrained maximizer is conditionally disposed to co-operate in ways that, followed by all, would yield nearly optimal and fair outcomes, and does co-operate in such ways when she may actually expect to benefit" (ibid., p. 177).

Notice also, as David Schmidtz has pointed out to me, that although Gauthier here describes his argument as showing that CM maximizes expected utility, it is more properly described as showing that CM is the dominant strategy (in cases of transparency).

[9] Ibid., p. 186.

5. Hence, if you choose SM, it makes no difference to my utility whether I choose CM or SM; whereas if you choose CM, I maximize my utility by choosing CM and inducing you to build.
6. So it is rational for me to choose CM.
7. If I am rational to choose CM, then it is rational for me to carry it out when the time comes (e.g., to build my lighthouse if you build yours).

Notice a highly significant feature of this argument: if successful, it entails that I should adopt CM and follow it when interacting with you, *even if you and I will only interact on this single occasion, and even if our interaction will have no effect on my future opportunities to cooperate with other individuals.* Gauthier rejects solutions to Prisoner's Dilemmas that rely on iterated occasions for cooperation, and takes himself to have shown that it is rational to cooperate even in the single-interaction case.[10]

Comment on premises 1 and 2

I have deliberately phrased premises 1 and 2 so as to reveal certain aspects of Gauthier's argument that his own language tends to obscure. Let me comment briefly upon these aspects.

First, I have worded premise 1 to bring out Gauthier's assumption that the choice of constrained maximization at t_1 will actually lead me to form the appropriate unconditional intention at t_2 and then perform the chosen act at t_3. For example, if I form the intention at t_2 of building my lighthouse, then I *will* build it at t_3. There is no possibility that I will fail to carry out what CM requires of me. We might call this assumption the *causal efficacy thesis:* the thesis that forming an intention to do A will cause the performance of A. Gauthier's calculation of my expected utility in choosing CM implicitly incorporates this assumption (for, if I could backslide on an intention to build my lighthouse, the expected utility of intending to build would not simply equal the expected utility of both of us building lighthouses). Indeed, given the truth of the transparency assumption, and the fact that, if you have chosen CM, you will only build your lighthouse if you know I will build mine, Gauthier needs the truth of the causal efficacy thesis to establish that my choosing CM will induce you to build and so maximize my utility. He also needs it to rule out the possibility of my adopting a strategy like the following one:

> KM: Forming the intention to build if you build, or not to build if you do not build; and then not building *whatever* you do.[11]

[10] Ibid., pp. 169–70.

[11] Some people object to KM on the ground that one could never intentionally adopt it, since that would involve simultaneously committing myself to (a) intending at t_1 to do A, and (b) intending at t_2 *not* to do A. It is claimed that one cannot, at least rationally,

Under conditions of translucency, adopting KM rather than CM might well maximize an agent's utility because it would enable him to mislead other agents about his future actions. However, KM is ruled out by the assumption of causal efficacy. But we should at least note the extreme strength of this assumption. I find it quite implausible to assume that any intention of mine *inevitably* causes my subsequent carrying out of that intention: some do, but some do not. Upgrading the *kind* of mental state I form (to a commitment or resolution) does not change this fact. Of course, we often change our minds when we acquire new information, or when we adopt new values. Gauthier wants to set such cases aside.[12] But even in cases where none of these factors matter, it is implausible to suppose our commitments always compel our future acts – especially in the kind of case in question, where considerations of utility press the agent to change her mind when the time comes. Indeed, if the causal efficacy thesis were true, it is hard to see how Prisoner's Dilemmas could have been the deep historical problem for social cooperation that they have been. However, in order to turn to other aspects of the argument, I will nonetheless provisionally grant Gauthier the truth of the thesis.[13]

The second point to notice is that I have stated the transparency assumption (in premise 2) to bring out the fact that it incorporates not merely the assumption that each person will be aware of the other's choice, but also the assumption that each person will know *that the other will actually carry out that choice.* Gauthier himself tends to discuss the transparency assumption in terms that suggest it merely implies that each person is aware of the other person's current mental state, such

commit oneself to this sort of inconsistent intention formation. Even if this is true, at most it would show that KM could not be recommended to agents by a *decision guide*, that is, a prescription to be used in their actual decision making. It would not show that KM fails to be the best disposition to actually have (even if one could not rationally bring it about that one has this disposition). If the causal efficacy thesis were not true, then KM might well be the best disposition for an agent to have in a Prisoner's Dilemma under conditions of translucency.

[12] David Gauthier, "Afterthoughts," in *The Security Gamble: Deterrence Dilemmas in the Nuclear Age*, edited by Douglas MacLean (Totowa, NJ: Rowman and Allanheld, 1984), p. 159.

[13] Would it raise difficulties for Gauthier's argument if my forming an intention to do *A* only caused me to do *A sometimes*, say, 60% of the time? Suppose you and I are both CMs, and I form the intention to build – but in this particular case, I will not carry through. Since I am transparent, you will detect my future failure to build, and so will form the intention not to build, and will not build. But since you are also transparent to me, I will detect your intention not to build, and so, by CM, should form the intention not to build myself. But by hypothesis, I have formed the intention to *build*. It appears as though failure of the causal efficacy thesis would be inconsistent with Gauthier's other assumptions.

As David Schmidtz points out to me, there is a related problem here, which is that Gauthier makes no provision for the possibility that either prospective partner might fail to carry out the intention to cooperate, not because of any change of mind, but rather because circumstances conspire to make cooperation impossible (e.g., I intend to build my lighthouse, but then am unable to purchase the requisite building materials). Part of this problem could be evaded by stating the constrained-maximization intention as a *doubly* conditional intention, namely, an intention to cooperate, if others do, and if the world permits. Such a restatement would mean that potential partners to whom one was transparent would have to know, not just one's intentions and future trustworthiness, but also whether the world would permit cooperation. But epistemic relations between potential partners would remain complex, and perhaps involve the kind of inconsistencies just described, in any case where the world will not permit cooperation.

It is worth pointing out that Gauthier's adoption of the causal efficacy thesis appears to raise a special problem for him. In his discussion of deterrence cases, he explicitly rules out cases in which the agent is *im*perfectly rational, unable fully to control her behavior in terms of her considered preferences. As an instance of such a case, he describes an agent whose cool preference not to retaliate will be overcome by her feelings of anger, rage, or panic at the moment action is called for. Gauthier stipulates that he only wants cases in which the person can control her behavior at the time when the choice to retaliate or not must be made. Second, Gauthier rules out cases in which the person, in expressing her intention, delegates her power to choose, by arranging that some other person or some preprogrammed device, capable of ignoring her preferences, will ensue that if the threatener strikes, retaliation will ensue. (Gauthier, "Deterrence, Maximization, and Rationality," in *The Security Gamble*, edited by Douglas MacLean, pp. 100–5.) Thus, we are explicitly confined to cases in which, at the time of action, the agent is rational and in control of what she chooses and does. Clearly, Gauthier wants the same conditions to hold in compliance cases as in deterrence ones.

These restrictions may appear to rule out the causal efficacy thesis, because it may appear that if a person is *caused* to do *A*, then she is not exercising rational choice to do *A* at the time of the action. However, my view, and I suspect Gauthier's as well, is that the causal efficacy thesis is *not* ruled out by the requirement that the person choose rationally at the time she decides whether or not to carry out her intention. (Notice a potentially confusing switch in terminology here: "choose rationally" does not mean, as it usually does in Gauthier's and my texts, "perform the best act"; rather, it means something like "make a reasoned choice given one's evidence.") Since I believe causation is compatible with rational choice, all this requirement entails is that the causal chain initiated by her forming the intention to do *A* must operate by causing her *to choose rationally to perform A*. It must cause her, for example, to deliberate rationally in selecting her alternative. Thus, forming the intention to build a lighthouse must cause her to believe that building a lighthouse is best. Since Gauthier is not interested in misinformed or irrational agents, this belief must be true. But we know that it would not be best for her to build a lighthouse unless she has previously formed the intention of doing so (since building a lighthouse, by itself, merely *de*creases her utility). Thus, forming the intention to build must not only cause her to believe retaliation would be best, it must *in addition make* it the case that building *is* best. So far, however, we have not established that forming this intention would make it best to build. Now, to support the final conclusion of his argument, Gauthier invokes a principle that asserts that if it is rational (best) to form an intention, then it is rational (best) to carry it out. (See the discussion in the fifth section.) Thus, if we could establish that it is rational for the agent to form the intention to build, then we could derive the needed conclusion here, namely, that forming the intention to build does indeed make it best to build. However, we cannot at this point in the argument help ourselves to the assumption that it is rational to form the intention to build, since this is precisely what we are trying to prove. We cannot argue that the intention is rational because it would cause the intended act to be performed, and then turn around and argue that the act itself is best because the earlier intention to perform it was rational. Given Gauthier's desired constraints, the intention is only rational if it would cause the intended act to be performed *rationally*, that is, if it would make the intended act best. And we cannot prove this until we have proved the intention is rational. There seems to be an unavoidable circularity in the argument here.

as the other person's current choice or intention. For example, he states that "A person's expectations about how others will interact with him depend strictly on his own choice of disposition only if that *choice* is known by the others."[14] However, it is clear that Gauthier both needs and relies on the much stronger assumption that others not only know the person's current choice, but also predict infallibly whether or not the person will carry out that choice. For example, Gauthier defines "transparency" as a state in which "each is directly aware . . . whether he is interacting with straightforward or constrained maximizers."[15] Since, as we have just seen, a constrained maximizer is one who *will* carry out her intentions, it follows that awareness that you are a constrained maximizer involves the true belief, not only that you are forming certain conditional intentions, but also that you will carry one of them out when the condition is satisfied. If I did not have this infallible ability to predict your future acts, Gauthier could not argue (as he does) that I maximize my expected utility by choosing CM, since I would sometimes be cheated by defecting partners. Correspondingly, Gauthier's translucency assumption is an assumption that agents' abilities to detect each others' intentions *and* to predict future fulfillment of those intentions surpass mere guesswork. My point here is not to reject these assumptions, but rather to indicate clearly how strong they are, and at least to raise doubts about their credibility.

Does constrained maximization maximize expected utility?

Now let us turn to premise 3, which spells out what the consequences will be if I choose CM as opposed to SM. Here the argument runs into trouble.

The first major problem is that Gauthier mistakenly assumes both that (a) my only options are CM and SM, and (b) my partner's only options are CM and SM. He then argues that CM is my best response to the possibility of these choices on the part of my partner. Both these assumptions seem false – there is no reason to suppose that our options are limited in this fashion. Since discussion of what additional options might be rational for me to choose is complicated, I restrict myself here to pointing out that my partner has a variety of alternatives, and that CM is not my best response to many of these.

I may, for example, face a partner who has chosen what we may call "unconditional cooperation (UC)":[16]

> UC: Building one's lighthouse whatever one's partner does.

[14] Gauthier, *Morals by Agreement*, p. 173, my emphasis.

[15] Ibid., pp. 173–4.

[16] I am grateful to David Schmidtz for pointing out to me the relevance of this alternative option.

Or I may face someone who has chosen "radical cooperation (RC)":

> RC: Building one's lighthouse if and only if one's partner has chosen unconditional cooperation.

How would someone who has chosen CM fare against a partner who has chosen UC? It is obvious that the best choice is SM rather than CM, since the straightforward maximizer will always obtain −500 and the constrained maximizer will only obtain −600. Similarly, against a partner who has chosen RC, one will maximize utility by choosing UC rather than CM, for in this way one will achieve a constant return of −600 rather than −1,000. Of course, a partner who chooses UC or RC may not be rational, since this choice will not fare well against all possible configurations of partners and their choices (although it is certainly arguable that someone choosing UC is *moral*). But CM is hardly an attractive option if it only succeeds against perfectly rational partners. Since many people are in fact irrational, an acceptable option must work against all comers, be they rational or not.[17] Hence Gauthier's argument that CM maximizes utility is significantly incomplete, since it does not take account of cases in which the chooser's partner will select some other option (perhaps an irrational one) besides either CM or SM. To establish whether CM is superior, we would have to know how many potential partners have chosen SM, CM, UC, RC, and all the other options that could be defined; how transparent or translucent these individuals are; what the stakes are in possible interactions with them; and so forth. Gauthier has not performed this task for us, and we can have no advance reason to suppose that CM will emerge ever or often as the best policy.

But even if we restrict ourselves to partners who choose either CM or SM, Gauthier's argument is less compelling than he realizes. Let us look at his first claim:

[17] The importance of establishing the rationality of a strategy against irrational opponents has been recognized since von Neumann and Morgenstarn concluded that ". . . the rules of rational behavior must provide definitely for the possibility of irrational conduct on the part of others." John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton: Princeton University Press, 1953). p. 32, as quoted in Brian Skyrms, *The Dynamics of Rational Deliberation* (Cambridge, MA: Harvard University Press, in press), chap. 6.

Even Gauthier, who argues that CM rather than SM is rational, implicitly concedes that CM must succeed against irrational dispositions, of which (in his view) SM is one.

In "Contractarianism and Moral Skepticism" (Chapter 13 in this volume), David Copp describes the contractarian strategy as arguing that there is a set of moral requirements such that a population of fully rational people would agree to comply with them. But if this argument assumes that each person may presuppose the rationality of everyone else in agreeing to comply with these requirements, it idealizes our actual situation beyond any usefulness.

If I choose CM, then
(a) If you choose CM, you will build and I will build.

Why, precisely, is it supposed to be the case that if I choose CM, then if you choose CM as well, you will build and I will build? Each of us is transparent, so each of us knows the other's choice, plus the fact that the other will carry out that choice. But how does each of us make the relevant choice?

There are actually *two* crucial choices that each of us must make here: the choice between CM and SM, and the choice (if CM is selected) of which unconditional intention to form, the intention to build or the intention not to build. Let's simplify matters by assuming that you have already chosen CM, and, of course, I know this. What should I predict you will do if I choose CM? Gauthier seems to assume that I may simply predict that you will form the intention to build your lighthouse, and then build it. But matters are much more complex than this. Given your choice of CM, what I *can* predict is that you will form the intention to build and carry it out if and only if you predict that I will form the unconditional intention to build and then carry out this intention. But how can I assume you can make this prediction about me? Your position vis-à-vis me is perfectly parallel to my position vis-à-vis you. Even if you know that I choose CM, all you can infer from this is that I will form the intention to build and carry it out *if and only if* I predict you will build. You cannot infer that I will intend and build, simpliciter. But if you cannot infer this, then you will not build. And I will not form the intention to build and then carry it out unless I believe you will build. Neither of us, knowing the other has chosen CM, has sufficient information to predict on that basis what the other will do, so neither of us can decide which intention to form and act to carry out.

What we have just seen is that if "transparency" is interpreted as Gauthier officially interprets it, namely, as awareness on the part of another agent whether he is interacting with a constrained or straightforward maximizer, then transparency is not sufficient to allow two transparent constrained maximizers to predict what the other will do, or to choose any intention or action themselves, since that choice depends on their making this prediction. We might try to avoid this by allowing Gauthier a *maximum* notion of transparency, namely, the assumption that each agent just directly knows (or truly believes) what (unconditional) intention the other agent forms, and knows that the other will carry out that intention. But even this will not solve the problem. To see this, all we need to do is notice that it is perfectly consistent with our both being constrained maximizers that we both form the intention *not* to build, and then carry out that intention. (For being a CM *may* involve forming the intention not to build if you predict the other agent will not build, and then not building yourself; if both of us do this, then we both still qualify as choosing and carrying out CM.) Thus, it is *not* possible, contrary to Gauthier's argument, to show that if we both choose CM, *and we are both transparent to each other in this* maximum sense, then we will both form the intention to build and then carry out those intentions. We might instead both form and act on the intention *not* to build. But if it is false that two constrained maximizers will necessarily cooperate, it is false that my choosing CM when you have chosen CM necessarily produces greater utility for me than my choosing SM: they might produce the very same utility – the utility I receive if neither of us builds a lighthouse.

It might be hoped that a reformulation of constrained maximization could be found according to which it would necessarily be the case that if two transparent partners both choose CM, then they will both build. Unfortunately, such reformulations do not seem to be available. Consider the following simple candidate:

> CM′. Forming the intention to build, and carrying out that intention.

Two agents, both of who adopt CM′, will necessarily both build their lighthouses. But adopting CM′ leaves each of them vulnerable to straightforward maximizers, since it directs the agent to build regardless of what her partner does. To avoid this, we need to incorporate the kind of clause adopted by Gauthier, which dictates adopting the same intention as one's partner; and to secure joint building, rather than joint nonbuilding, we need an additional clause that provides pressure, so to speak, in the direction of building. The following option might be thought to solve the problem:

> CM″. (i) Forming the intention to build if one's partner will build, or forming the intention not to build if one's partner will not build; and
> (ii) carrying out whichever intention one forms; and
> (iii) forming the intention to build and actually building if one's partner forms the intention to comply with (i) and (ii).

It may appear that CM″ secures the result Gauthier wants, since if both you and I adopt CM″, then each of us forms the intention to comply with clauses (i) and (ii), and, hence, each of us must comply with clause (iii) – so we both build our lighthouses. However, CM″ must be rejected, for it does not prescribe a consistent set of intentions. For suppose you have adopted clauses (i) and (ii) (but not clause (iii)), and you also form the intention not to build (as we have seen, nothing in clauses (i) and (ii) rules this out). Then, by CM″, I must *both* form the intention to build (since that is required by clause (iii)), and *also* form the intention not to build (since that is required by clause (i)). But these intentions, and their attendant actions, are inconsistent. In general, every reformulation of CM I have inspected is vulnerable to some problem or another: either it does not guarantee joint compliance, or it leaves the agent vulnerable to exploitation by a partner who does not adopt CM, or it delivers inconsistent prescriptions. Unless a successful version of CM is forthcoming, we must conclude joint adoption of constrained maximization cannot secure joint compliance as Gauthier's argument assumes.[18]

[18] In "Closing the Compliance Dilemma: Why It's Rational to be Moral in a Lamarckian World" (Chapter 16 in this volume), Peter Danielson recognizes this difficulty, and formulates a version of CM that he claims avoids the problem. He introduces the notion of a "metastrategy," that is, "a function that takes each of the other player's choices (or metastrategies) into a choice." He then introduces the metastrategy CC (intended to be roughly similar to Gauthier's CM), which is defined as "CC = UC → C; MAX → D; CC → C," (where UC and MAX are other metastrategies). But *this* definition is clearly illegitimate, since it refers to the very concept being defined (CC itself). Danielson states that CC can be defined "extensionally" in terms of the game matrix, but this cannot be correct, since he explicitly recognizes (his Figures 16-1 and 16-2) that a first-order game (in which the agents' options are ordinary actions) may have the same matrix as a second-order game (in which the agents' options are metastrategies). Thus, game matrices alone cannot define metastrategies, since such matrices do not distinguish them from ordinary actions or choices. Indeed, since a "metastrategy" is *defined* as a strategy from another player's choice or metastrategy to a choice, it is difficult to see how a metastrategy could be individuated (much less implemented, as Danielson himself realizes, p. 298), without some reference to the other player's choice or metastrategy. In other work, Danielson attempts to solve this problem by use of a quotational device that I have not had the opportunity to examine.

In "Gauthier's Theory of Morals by Agreement" (*Philosophical Quarterly* 38 (1988): 343–64), Richmond Campbell suggests the following formulation of CM to get around this problem: "In a choice situation involving strategic interaction a person has the CM disposition iff: (1) she has property R and (2) she will co-operate with other agents interacting with her iff she believes that each of them has property R." Property R is any property that SMs won't have, such as the property of being ready to reciprocate cooperation when making the second move in *sequential* Prisoner's Dilemmas. It is true that SMs won't have the property Campbell describes, so that someone who adopts CM will not be exploited by SMs. However, there will be other defecting agents who *would* reciprocate cooperation in sequential Prisoner's Dilemmas, but who would *not* cooperate in simultaneous Prisoner's Dilemmas; an agent adopting Campbell's version of CM would be vulnerable to exploitation by these agents. SM agents do not exhaust the list of possible noncooperative agents that CM agents must be protected against. We cannot, of course, without problematic circularity, simply define a CM agent as one who cooperates with all and only other CM agents (or ones she believes to be CM agents). There is a deep problem here about whether the kind of relation Gauthier needs between CM partners is incoherent: it may be that his conditions will only be met if the decision of each partner *causes* the decision of the other partner – a type of causal interaction that cannot be countenanced.

In *Utilitarianism and Co-operation* (Oxford: Clarendon Press, 1980), Donald Regan discusses analogous problems of coordination that might face act utilitarians. In Chapters 8–10, he introduces a new version of utilitarianism that, he argues, enables agents who follow it to protect themselves from "exploitation" by others, and yet enables them as a group to secure the greatest utility available to them as a group. The daunting complexity of this new theory leaves me uncertain as to its success; but readers interested in pursuing this problem for Gauthier would be well repaid to study Regan's contribution.

On Gauthier's original presentation of his argument, I never do worse if I choose CM than if I choose SM, and in one case out of four possible types of cases, I will do better. On my construal of the argument (assuming my partner chooses either SM or CM), I never do worse if I choose CM, and in one case (where you also choose CM) out of four, I *may* do better. This is getting to be pretty slim pickings as support for the rationality of choosing CM. It becomes even slimmer when we move from the strong transparency assumption to the more realistic translucency assumption Gauthier himself uses. Under the transparency assumption, CM is a safe choice, since I can never be exploited by straightforward maximizers: I can always recognize them in advance and protect myself. But under the translucency assumption, CM is no longer safe: it exposes me to exploitation by undetected straightforward maximizers. Both I and potential partners can resort to the *pretense* of being a constrained maximizer, and then prey on those who are deceived. The issue then is whether the gain possibly available through adopting CM is outweighed by the risk of exploitation that it creates.

Gauthier argues that under the translucency assumption, it is rational to choose CM only if the ratio between the probability that an interaction involving CMs will result in cooperation and the probability that an interaction involving CMs and SMs will involve exploitation and defection is greater than the ratio between the gain from defection and the gain through cooperation.[19] In arguing that this will often happen, Gauthier assumes that two constrained maximizers who successfully identify each other will always cooperate. What we have seen is that they will *not* always cooperate. Hence, the probability that an interaction involving CMs will result in cooperation is lower than Gauthier assumes (although we cannot say how much lower). Therefore, it takes a correspondingly *lower* ratio between the gain from defection and the gain through cooperation for it to be rational to choose CM. The less people stand to gain from defection, the more likely it is that CM is rational; but the more people stand to gain from defection, the less likely it is that CM is rational. The rationality of choosing CM is not undermined for *every* case by the phenomenon I have described. What I have shown, however, is that the number of cases in which it is rational to choose CM is smaller than Gauthier supposes, and it may be significantly smaller. And, as we saw at the beginning of this section, it may not be rational at all if one's potential partners adopt options such as UC or RC rather than CM or SM. All this may dramatically reduce the number of occasions on which it is rational to dispose oneself to act morally, and so may considerably shrink the scope of the justification for moral action that Gauthier is trying to provide.

[19] Gauthier, *Morals by Agreement*, p. 176.

The alleged rationality of carrying out rational intentions

In the preceding section, I argued that Gauthier succeeds in showing the rationality of choosing constrained maximization in fewer cases than he realizes. Let us now turn to the question of whether it is really rational for me to carry out constrained maximization even in the cases where it is rational to choose it in advance. Line 7 of Gauthier's argument asserts that it is: that, for example, if I rationally choose CM and form the intention to build my lighthouse, then it is rational for me to actually build the lighthouse when the time comes.[20]

Why does Gauthier think it is rational for me to carry out my chosen policy? He admits that doing so is not rational in the usual sense of maximizing utility. It appears that what he relies on here is a general principle, which we may label the "rationality of perseverance" principle (RPP):

RPP: If it is rational for an agent to form the intention to do A, then it is rational for the agent to actually do A when the time comes (assuming the agent acquires no new information, and has not altered her values).

Gauthier asserts, for example, "If it is rational for me to adopt an intention to do *x* in circumstances *c*, and if *c* comes about . . . then it is rational for me to carry out *x*"; and also, "If [a person's] dispositions to choose are rational, then surely her choices are also rational."[21]

Unfortunately, and surprisingly, Gauthier offers no positive argument in favor of the rationality of perseverance principle. The only argument he appears to provide is the concession that if the agent will suffer from some future weakness or imperfection (such as weakness of will), then it will *not* be rational for her to persevere in carrying out an intention.[22]

But, of course, this does not provide positive evidence that rationality requires perfect agents to persevere.

Indeed, it is difficult to know how one could argue for the rationality of perseverance. On the face of it, intentions and the intended acts are two distinct events, sometimes having different features and different consequences. Hence, the appropriateness of an intention appears to imply nothing conclusive about the appropriateness of the intended act. I may, for example, have promised to form a certain intention, but not promised to carry it out; in such a case, I am obliged to form the intention, but not obliged to carry it out, especially if my doing so would be undesirable.[23]

The natural move here for a defender of perseverance would be to appeal to intuitions about rationality in various cases. Unfortunately, such appeals do not appear to support the principle. In standard cases, where the consequences of the intention are the same as those of the intended act, where the agent is rational to form an intention at t_1 to perform *A* rather than *B* at t_2, and where she receives no new information and does not change her values by t_2, we all agree that she then would be rational at t_2 to carry out her intention by performing *A*. But we cannot use this fact to support the claim that the existence of her prior intention to do *A* *makes it* rational to do *A*, since it would be rational for her to do *A* whether or not she had ever formed that prior intention. She has all the same reasons to do *A* at t_2 that she had at t_1; if it was reasonable to choose *A* over *B* at the earlier time, of course, it is reasonable to choose it now. Someone attempting to support the rationality of perseverance principle by appeal to this kind of case would have to resort to a different scenario: a scenario involving two possible worlds, identical to each other except that in the first world, the agent forms a rational prior intention to do *A*, whereas in the second world, the agent forms no prior intention. It would then have to be claimed that at the time of action, the first agent has *more* reason to do *A* than the second agent. I, myself, have no inclination to think this: the two agents appear to me to have precisely equal reason to do *A*.

Gauthier needs the strong interpretation of the RPP, according to which the rationality of the earlier intention *makes* the later action rational. Otherwise, in the kinds of cases with which we are specifically concerned, that is, cases where the agent's reasons for forming the

[20] Part of this claim is the assertion that if I rationally choose CM, and, seeing that you will not build your lighthouse, I form the intention not to build *my* lighthouse, then it is rational for me to carry out this intention as well. But no one disputes that carrying out this particular intention is rational.

[21] Gauthier, "Afterthoughts," p. 159, and *Morals by Agreement*, p. 186.

[22] Gauthier, *Morals by Agreement*, pp. 184–6. In discussing what it would be rational to do for agents who are subject to weaknesses or imperfections, Gauthier states that such rationality "constitute[s] a second-best rationality," and denies that any lesson can be drawn from this about the dispositions and choices that are rational for the perfect actor (pp. 185–6). This seems to be a mistake. It is rational, simpliciter, to do the best one can in the face, so to speak, of the materials that have been given to one. An imperfect agent, one afflicted with future weakness of will, does the best she can in the face of *this* material. The material is imperfect, but her rationality in dealing with this material is not imperfect or "second best." Similarly, an agent facing a threatener, or a straightforward maximizer, does the best she can in the fact of this material. The agents she faces are also imperfect (in Gauthier's view), but this does not show that her rationality in dealing with these agents is imperfect. I can see no principled differences between doing the best one can vis-à-vis one's own future imperfections and doing the best one can vis-à-vis other agents' imperfections.

[23] Since this essay was written, I have discovered that similar arguments to those in this paragraph (and a case similar to my subsequent telepathic terrorist case) have already been presented by Gregory Kavka in "Responses to the Paradox of Deterrence," in *The Security Gamble: Deterrence Dilemmas in the Nuclear Age*, edited by Douglas MacLean (Totowa, NJ: Rowman & Allanheld, 1984), pp. 155–9.

A second independent and insightful discussion of what I call RPP (keyed to another article of Kavka's) is contained in Michael Bratman, *Intention, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press, 1987), pp. 101–6.

intention to do A are independent of his reasons for actually doing A, the action would have to be judged irrational according to the normal utility-maximizing criterion of rationality. This presumption of irrationality could only be overturned by the strong interpretation of the RPP. But it is precisely in such cases that we feel the greatest conviction that the agent's reasons for forming the intention to do A seem not to carry over *at all* as reasons to do A itself. Consider the case of a government official negotiating with a terrorist. The terrorist, an infallible mind reader, threatens to blow up a planeload of innocent people unless the official forms the mental intention of releasing the terrorist's imprisoned comrades. When, and only when, the official forms this intention (a psychological event the terrorist will detect telepathically), the bomb will be disarmed. Clearly, under these circumstances, it would be rational for the official to form the intention of releasing the terrorist's comrades. Once he has formed the required intention and the terrorist has permanently disarmed his bomb, it seems equally clear that it would be rational for the official to change his mind and *not* release the comrades. Indeed, we would view a terrorist who merely demanded the formation of intentions as extremely stupid. Here, then, is a clear-cut case where it is rational to form an intention, and yet irrational to carry it out, contrary to the implications of the rationality of perseverance principle.[24] Gauthier might argue that his acceptance of the causal efficacy thesis shows that it is causally impossible for the official both to form the intention to release the comrades and then to change his mind. But this does not show that actually releasing the comrades is *rational,* that is, the best act. Gauthier assumes throughout that the agent has genuine alternatives available to him at t_2, that is, that he *could* fail to release the comrades. And if he could fail to release them, it is open to us to assess this act as the best one.

It may fix this conclusion even more firmly in our minds to notice that Gauthier needs an extremely strong perseverance principle: not just the assertion that prior intentions create *some* reason to carry them out, but rather the assertion that prior intentions create *conclusive* reason to carry them out. For if the reason to comply created by one's earlier intention is not conclusive, there will always be the danger that the disutility of compliance will outweigh the reason in favor of compliance created by the earlier intention, and defection will turn out to be rational after all. The complete counterintuitiveness of such a strong perseverance principle can be seen if we merely raise the stakes at issue in a given case. Consider a case in which a telepathic burglar threatens to steal all my household valuables. I know that if I form the intention of blowing up the house with the burglar and myself inside, it is nearly certain that he will be deterred. I form this intention, but unfortunately he is not deterred. According to the strong perseverance principle, it is now rational for me to blow up the house and kill myself, merely because I previously formed the intention of doing so under these circumstances. But no one, I think, would want to agree to this. Gauthier's guiding hope in his project is to provide a foundation for moral norms on norms of rationality, because the latter's binding status is much more easily accepted. But if it turns out that the norms of rationality include such principles as perseverance, it appears they may demand greater self-sacrifice from us than even morality typically does, and will be even more difficult to accept. If morality requires rationality of perseverance as its foundation, we are probably worse off than we are with a morality that has no foundation in rationality.

It is also worth noting that nothing in Gauthier's argument for the rationality of choosing constrained maximization really turns on the fact that the means by which I effect my subsequent compliance is a choice to form a certain *intention.* We could restate the entire argument, just as effectively, to show that it would maximize my utility to perform *any* present act that would bring about my subsequent choosing to build a lighthouse. For example, suppose that tonight I open a magazine to an article I want to read tomorrow. Tomorrow I read the article; its mentioning lighthouses triggers the idea that I might build one, and that is precisely what I do. Suppose also that I am transparent to you, in the sense that you detect my opening the magazine tonight and accurately predict that it will stimulate me tomorrow to build a lighthouse. You are a (suitably defined) constrained maximizer. If Gauthier's argument were correct, my opening the magazine in these circumstances would induce you to build your lighthouse as well, and we would have evaded the Prisoner's Dilemma. Hence, since it would (in Gauthier's argument) maximize utility for me to open the magazine, rationality requires that I do so. But, when the time came for me to build the lighthouse, Gauthier would be unable to argue that building the lighthouse is also rational for me, since he can invoke no prior intention on my part to establish its rationality. Since there is no difference between the utility I can secure by forming an intention and the utility I can secure by opening the magazine, there is no way for Gauthier to argue that I must form the intention rather than open the magazine. Agents who bring about their own compliance with moral rules by such means as opening magazines cannot be said to be rational when they do comply.[25]

[24] Gauthier regards deterrence cases and compliance cases as completely parallel, so he would not dismiss such a case as irrelevant to his argument about compliance with morality. (See *Morals by Agreement*, pp. 184–7.)

[25] It is clear that in opening the magazine, I have no intention to build a lighthouse. We might investigate whether I could *deliberately* secure my building the lighthouse by performing some act today that would bring about my building the lighthouse tomorrow, without at the time of the act intending to build a lighthouse (it might be claimed,

We should note the flip side of Gauthier's reliance on intentions to make subsequent compliance rational. Imagine a case in which one's partner has performed the cooperative act, even though one did not oneself adopt constrained maximization, or form the intention to cooperate. In such a case, Gauthier must say that it is not rational, and hence not moral, for one to cooperate, since no prior intention to cooperate makes it rational now. (And, since belatedly forming such an intention would have no effect on one's partner's action, there is no reason to do so, even if such an intention, if rational, would make one's cooperating rational.) This seems a clear case in which the recommendations generated by Gauthier's system diverge sharply from those of ordinary morality, which would recommend cooperation even if one had not originally intended to cooperate.

I conclude that the rationality of perseverance principle is false.[26] If this is correct, then Gauthier has not shown that it is rational to carry out constrained maximization, even in those cases where it is rational to adopt it. It follows from this that he has not shown that it is rational to abide with moral rules once accepted, even if it is rational to accept them. Moreover, even if the rationality of perseverance principle were true, it would not apply to cases in which an agent induces himself to comply with moral precepts by some means (such as opening a magazine) that do not involve forming an intention to perform the act of compliance. Any agent who maximizes his utility by such means cannot be said to be rational when he then complies. Whether his act of compliance is brought about by a utility-maximizing intention or a utility-maximizing act of another sort, the compliance itself is not rational.

The derivation of morality from rationality

I have argued that Gauthier has not shown that it maximizes expected utility in a significant number of cases for a person to adopt constrained maximization. I have further argued that he has not shown, because it is false, that it is rational for a person to act according to constrained maximization even in the cases where it was rational to adopt it. Now I want to take up a harder and more important question. Suppose my previous arguments were wrong, and Gauthier had indeed shown that it is rational both to adopt constrained maximization in many cases and then to carry it out. Would the success of this argument show, as Gauthier believes it does, that morality is founded on rationality, and hence that rationality provides a "justificatory framework for moral behavior and principles"?[27] If the success of Gauthier's argument *would* provide a rational justification for morality, then we would be well repaid to tinker with the details of his argument in an attempt to salvage it from my previous criticisms. But if success would not provide such a justification, then such tinkering has little or no point.

I shall assume that the point of providing morality with a justificatory framework is to defend it against the moral skeptic, who believes all moral statements are false, or meaningless, or at any rate without epistemological justification. We may characterize what Gauthier has done as arguing that individual rationality, or self-interest, requires a person to dispose herself to perform certain cooperative acts, and then actually to perform those acts when the time comes. Suppose we assume that the acts in question are precisely the same ones that morality requires. Still, the success of this argument would not show that *morality* has been provided with a justification. It would show that we have self-interested reasons to do what morality, *if it were true* (or correct), would demand – but it would not show that morality *is* true (or correct). Such an argument would merely show an interesting coincidence between the purported claims of morality and the real claims of self-interest.[28] In

for example, that doing *A* with the intention of bringing about my doing *B* shows that I *do* have the intention of doing *B* at the time I do *A*). I am not at all sure this latter claim is correct, but I shall not try to discuss it here. The important point for us here is that my act of opening the magazine is rational, that is, utility maximizing, whether or not *I believe* it will have the desired effect. And if I open the magazine in ignorance of its effect, I can hardly be said to have the intention of bringing about that effect.

Gauthier might try to assert (and there is some textual evidence that he believes) an analogue of the perserverance principle: that if it is best (rational) for me to do act *A* (partly because it will bring about later act *B*), then when the time comes, it must be best (rational) for me to do *B* as well. Clearly, this is incorrect. Suppose a business student has only two choices: to accept a job with firm 1, or to accept a job with firm 2. If she accepts the job with firm 1, she will subsequently cheat a client out of $100, be caught, and sentenced to spend a month in jail. If she accepts the job with firm 2, she will subsequently cheat a client out of $100,000, be caught, and sentenced to spend ten years in jail. Given the prospects, it would be best for her to accept the job with firm 1 – and best partly because this job will result in her stealing (merely) $100. But it hardly follows that it would then be best (or rational) for her to cheat the client out of $100. She may be destined to do this by her earlier choice, but we can still criticize her action. (See above, footnote 20, for a discussion of whether such cases are sufficiently parallel to the ones in which Gauthier is interested.) In cases such as this, one might question whether act *B* (stealing $100) is part of what makes act *A* (accepting the job with firm 1) best – it would be more intuitive to say that act *A* is best *despite* its leading to act *B*. But the analogue to the perserverance principle just articulated is only significant if it applies to subsequent acts (such as stealing the $100) that are not rational or best taken in themselves.

[26] In "Contractarianism and Moral Skepticism" (Chapter 13 in this volume), David Copp also argues (in the third section) that actual compliance is irrational even if the CM disposition is rational (except in the sense that the CMs preferences have changed, in virtue of her adoption of CM, so that compliance necessarily becomes the utility-maximizing act).

[27] Gauthier, *Morals by Agreement*, p. 2.

[28] I take this to be one of the points of David Coop's discussion in "Contractarianism and Moral Skepticism" (Chapter 13 in this volume). Copp and I initiated our lines of inquiry into the force of Gauthier's argument independently; reading earlier versions of his stimulating discussion forced me to clarify for myself what I thought Gauthier was really trying to do in connecting rationality and morality.

order for Gauthier to answer the moral skeptic, he must do something more. But precisely what?

One promising strategy for answering the moral skeptic would be to find premises that the skeptic must accept, and then to show that certain moral statements follow from those premises. I believe that Gauthier can be interpreted as trying to follow precisely this strategy. He puts forward certain premises that he believes everyone must accept and then argues that moral statements can be derived from them. In his case, the premises are not factual statements, or definitions of moral terms, but rather normative principles of individual rationality. If these principles cannot be rejected, or at any rate are more readily acceptable than morality itself, then if moral principles can be shown to follow from them, morality will have been provided with a suitable foundation. We need not follow the cognitivist in granting that the principles of individual rationality are *true*, but whatever variety of acceptability they do have will be fully inherited by the moral principles that can be shown to follow from them.[29] The moral skeptic will be hard put to reject such principles.

How precisely is the argument supposed to go? Gauthier posits the following two axioms of individual rationality (where RPP is our old friend, the rationality of perseverance, and RMX is a *qualified* version of he rationality of maximizing expected utility). He characterizes these as comprising a "weak and widely accepted conception of practical rationality."[30]

RPP . If it is rational for an agent to form the intention of doing *A*, then it is rational for the agent to actually do *A* when the time comes (assuming the agent acquires no new information, and has not altered her values).

RMX. It is rational for an agent to act so as to maximize her expected utility, unless doing so requires her to violate RPP.[31]

He then argues that these axioms entail the following principles:

1. Under circumstances C, it is rational for an agent to adopt constrained maximization and form the intention to cooperate.

[29] Some authors have tried to follow this strategy by deriving moral statements from definitions of moral terms. For example, if we agree what "is morally obligatory" just means "would maximize the general happiness," then we can hardly reject the truth of the statement "It is morally obligatory to maximize the general happiness." The weakness of this version of the strategy is that a skeptic can respond, "Why should anyone care about acts that maximize the general happiness?" Gauthier's version avoids this problem, because it *already starts* with normative principles that have motivational force.

[30] Gauthier, *Morals by Agreement*, p. 17.

[31] Ibid., pp. 43–4 and 182–7.

2. If the agent rationally forms the intention to cooperate, then it is rational for her to carry out this intention (assuming she has acquired no new information and has not altered her values).

The phrase "circumstances C" refers to any combination of empirical factors, involving the intentions of other agents, their degree of translucency, the possible gains and losses associated with cooperation and defection, etc., that makes it true in a given case that the agent would maximize her expected utility by adopting constrained maximization. Principle 1 follows straightforwardly from axiom RMX, if we allow this form of rationality to be applied to the formation of intentions (i.e., adoption of constrained maximization) as well as to external conduct. Principle 2 is a straightforward special case of axiom RPP.

Gauthier claims that the traditional conception of morality identifies any impartial constraint on self-interested behavior as moral.[32] On this account, Principle 1 is not a moral principle, since it in no way constrains self-interested behavior. To the contrary, it recommends adoption of constrained maximization precisely because it will maximize the agent's self-interest. Hence, if this argument provides a justification for morality, the whole weight falls on Principle 2. Principle 2 clearly does constrain self-interested behavior, since it requires the agent to cooperate even when doing so would fail to maximize her utility. Let us grant, for the sake of argument, that Principle 2 is also impartial in some suitable sense.

But does the fact that Principle 2 constitutes an impartial constraint on self-interested behavior show that it is a moral principle? Clearly not. Impartial constraint is *not* sufficient to show a principle is moral. Consider a rule of etiquette requiring thank-you notes to be handwritten rather than typed. Such a rule is certainly an impartial constraint, but it does not thereby qualify as a moral principle.[33] What more is required for morality? This is a difficult question, to which I will not try to supply a general answer. But it appears to me that we can characterize what is lacking in rules of etiquette as something like *appropriate deontic force.* It is by no means easy to say what the deontic force appropriate to morality

[32] Ibid., pp. 2, 4.

[33] Another example is the principle of malevolence, which prescribes any action maximizing the general unhappiness. This principle is both impartial and a constraint on self-interest (since it is often highly damaging to an agent's own interests to follow it), yet it hardly seems to be a moral principle. Other examples are supplied by club rules, legal codes, Mafia codes of honor, professional codes, administrative regulations, etc. Of course, some of the prescriptions stemming from these sources will require acts that are morally right, but it does not follow that all such prescriptions coincide with morality, or that any of them *in itself* constitutes a moral prescription. The difficulty with the principle of malevolence is not the one I cite in the text – inappropriate deontic force – but rather inappropriate content.

amounts to. But, borrowing from the traditional literature on the nature of morality, we might suggest the following three features:

I. Moral prescriptions are *overriding*. That is, they outweigh prescriptions from any other source when there is a conflict. In particular, they outweigh the recommendations of self-interest; or perhaps, more accurately, strong moral considerations outweigh weak considerations from any other normative sphere.
II. Moral prescriptions are *categorical*. That is, they hold independently of the agent's actual desires and aversions.
III. Moral prescriptions for action make it appropriate for agents to hold associated distinctive moral attitudes, such as guilt for personal derelictions, blame toward others who violate the prescriptions, a feeling that one is justified when one follows the prescription, etc.

The question before us is whether Principle 2 has the kind of deontic force required of genuine moral principles.[34] Is its recommendation overriding and categorical, and does it support attitudes of guilt and blame, etc.? I am not going to try to answer this question definitively. One can certainly argue that Principle 2 generates prescriptions that are both overriding and categorical. They are overriding because they always outweigh the recommendation of self-interest to maximize one's own utility, and they are categorical for the same reason: they tell the agent what to do regardless of her desires at the moment of action. (On the other hand, there is a sense in which they are neither overriding nor categorical, since these prescriptions only arise because of the agent's prior attempts to satisfy her desires and maximize her self-interest by adopting constrained maximization. This is not the kind of independence from desires that Kant, for example, had in mind.) It is far less clear that one can argue that prescriptions generated by Principle 2 appropriately support attitudes of blame, guilt, and so forth. I myself see no reason why an agent should feel *guilty* (or should *blame* others) for violating Principle 2, which is essentially a demand that one's action and intentions show a certain form of consistency. Inconsistency is not usually the object of blame and guilt.

But this is not the point I want to make here. What I want instead to point out is that whatever deontic force Principle 2 does possess, it inherits this force directly and solely from the axiom of rational perseverance, since it is simply a special case of that axiom. If that axiom has appropriate deontic force, then Principle 2 will as well, whereas, if that axiom fails to have the appropriate force, then Principle 2 will also fail. What this means is that Principle 2 only qualifies as a moral principle if axiom RPP qualifies as a moral principle. But if axiom RPP qualifies as a moral principle, then Gauthier has not succeeded in deriving morality from some nonmoral source.[35] Instead, he has derived morality from morality itself. This is no help against the moral skeptic. The only possible way in which such an argument could be construed as an effective response to the moral skeptic would be if the RPP, although moral, were somehow less questionable or more readily acceptable than other moral principles. Quite the opposite seems to be the case, as my arguments in the last section aimed at showing.

What we have discovered is that even if Gauthier had succeeded in establishing the truth of Principles 1 and 2, he would not have answered the skeptic by showing how morality can be derived from individual rationality. Either the principles he derives do not qualify as moral principles, because they lack the required deontic force, or if, on the other hand, they possess that force, it is only because a covert moral principle was smuggled into the axioms of individual rationality. No amount of tinkering with the details of Gauthier's argument will circumvent this problem.

This problem is a general one, extending beyond Gauthier's project to other possible attempts to derive morality from individual rationality. It appears to be precisely the powerful deontic force of morality that makes it suspect. It looks as though *any* attempt to derive morality in the manner I have outlined here faces the same dilemma as Gauthier's attempt: either the derived principles will be deontically too weak to qualify as genuinely moral or else the premises will be so strong that they already qualify as moral, and subject to the full blast of the skeptic's suspicion. If morality is to be defended against skepticism by an appeal to rationality, we need some better strategy than the one I have outlined here.

I conclude that the long-sought proof for the rationality of morality still eludes us, despite the hard and illuminating work Gauthier has devoted to the cause of finding it. It may elude us forever.[36]

[34] Gauthier's commitment to deriving principles with moral force is shown in the following passage: "A person is conceived as an independent center of activity, endeavoring to direct his capacities and resources to the fulfillment of his interests. He considers what he can do, but initially draws no distinction between what he may and may not do. How then does he come to acknowledge this distinction? How does a person come to recognize a moral dimension to choice . . . ?" (*Morals by Agreement*, p. 9).

[35] See ibid., pp. 5 and 17, for Gauthier's assertions that he aims to derive morality from a nonmoral source.

[36] I am grateful to Michael Bratman, David Copp, Peter Danielson, Alan Nelson, David Schmidtz, and George Smith for comments on earlier versions of this essay.