Published as Duncan MacIntosh, "Preference's Progress: Rational Self-Alteration and the Rationality of Morality", Dialogue: Canadian Philosophical Review, XXX (1991), pp. 3-32. See publisher’s version for exact footnoting and pagination. In the published version, the first footnote appears as an *, and all notes appear as endnotes.

Preference's Progress: Rational Self-Alteration and the Rationality of Morality[1]

1. Introduction

On the received theory of rational choice, (a) a choice is rational if it maximizes one's individual expected utility. However in the Prisoner's Dilemma (PD), by this standard, each agent should Defect, since each then maximizes no matter what the other does. But then both will Defect, doing poorly; better for each if both had Co-operated.

David Gauthier tries to escape this by conceiving rationality differently. On one reading of his proposal, (b) a choice is rational if it expresses a disposition adoption of which maximizes. Each agent should dispose himself to Co-operate with those like disposed, and then Co-operate with them. Such agents, he thinks, would both Co-operate, doing well.

In this paper, I do six things. First, I argue that Gauthian rationality does not really rationalize Co-operation, for it rationalizes reversion to a disposition to Defect after the other agent has chosen among actions. Second, I show that classical rationality can already rationalize Co-operation, for it justifies revising one's preferences so that one prefers to Co-operate with those with certain preferences; Co-operation with them is then justified as a straightforwardly maximizing choice, given one's new preferences. We learn then that it is a consequence of classical rationality that an action is rational just if it expresses rational preferences (ones which it is rational to have), and that a set of preferences, P, is one which it is rational to have just if no other set, P*, is such that having it would better advance maximization on P than would having P. Third, I consider whether a genuinely moral action can be a maximizing action, as it seems it must be able to be, if to be moral is to Co-operate; and if, as in my account, one Co-operates because one has rationally acquired a preference to Co-operate. Fourth, I detail the structure of the

[1] This paper expands, clarifies, refines, defends and considers some of the implications of, theses which I introduced in "Two Gauthiers?" Dialogue: Canadian Philosophical Review 28 (1989): 43-61. For this reason it might have been subtitled, "A Post-Script to `Two Gauthiers?'". For helpful criticism, my thanks to Robert Bright, Julia Colterjohn, Peter Danielson, Robert Martin, Victoria McGeer, Wayne Sumner, Terry Tomkow and especially Richmond Campbell, all of whom wrote me comments. I am also grateful to a careful and sympathetic copy editor, a patient production staff, and to Christopher Morris and two other (anonymous) referees who gave me very detailed and useful comments, and who asked more good questions than I will be able fully to deal with here. I thank Dalhousie University for the Isaac Walton Killam Postdoctoral Fellowship which funded early work on this paper.

preferences which it is rational to adopt for a PD, and explore the character of the solution in detail. Fifth, I answer objections from the nature of preferences and rationality. Finally, I reflect on the solution's generality. I seem to be arguing that one can only choose rationally by maximizing on one's preferences; one maximizes in altering one's preferences, then maximizes again in choosing Co-operation, given those new preferences.[2] But can one not, alternatively, rationally choose and rationally act on principles,[3] dispositions (e.g., those recommended by rational principles),[4] reflexes,

[2] I identify the problem to which this is a solution in my "Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem," The Southern Journal of Philosophy, 26 (1988): 399-425; briefly sketch it as a way of saving Gauthier in my "Two Gauthiers?"; and defend the intelligibility of revising one's preferences on pragmatic pretexts in my "Preference-Revision and the Paradoxes of Instrumental Rationality" (forthcoming, Canadian Journal of Philosophy). The latter, tracing from my "Retaliation Rationalized" (paper presented to the Canadian Philosophical Association, Learned Societies Meetings, May, 1988, Windsor, Ontario, now expanded and published as "Retaliation Rationalized: Gauthier's Solution to the Deterrence Dilemma," Pacific Philosophical Quarterly, Vol. 72, No. 1 (1991): 9-32), and my "Kavka Revisited: Some Paradoxes of Deterrence Dissolved" (unpublished manuscript, Dalhousie University, 1990), asks whether a harm-hater can rationally threaten nuclear retaliation to deter attack, and whether he can act on that threat if deterrence fails. This is congruent with the problems of the PD, since both involve a maximizing commitment to a non-maximizing action. E.g., see David Gauthier, "Deterrence, Maximization, and Rationality," Ethics, 94 (1984): 474-495; and "Afterthoughts," in The Security Gamble: Deterrence Dilemmas in the Nuclear Age, edited by Douglas Maclean (Totowa, NJ: Rowan and Allenheld, 1984), p. 159-161; Gregory Kavka, "Some Paradoxes of Deterrence," The Journal of Philosophy, 75 (1978): 285-302; "The Toxin Puzzle," Analysis, 43 (1983): 33-36; "Responses to the Paradox of Deterrence," in Maclean, ed., The Security Gamble, p. 155-159; David Lewis, "Devil's Bargains and the Real World," in Maclean, The Security Gamble, p. 141-154, and Mark Vorobej, "Gauthier on Deterrence," Dialogue: Canadian Philosophical Review, 25 (1986): 471-476.

[3] For a defence of this view, see Frederic Shick, Having Reasons: An Essay on Rationality and Sociality (Princeton, NJ: Princeton University Press, 1984).

[4] This construction can be placed on David Gauthier, Morals By Agreement (Oxford: Clarendon Press, 1986); and he explicitly defends it in his "In the Neighborhood of the Newcomb-Predictor (Reflections on Rationality)," Proceedings of the Aristotelian Society, 89 (1988/89): 179-194, and "Economic Man and the Rational Reasoner" (unpublished manuscript, University of Pittsburgh, 1987). He has also recommended this interpretation in correspondence. Peter Danielson has argued for a similar position in correspondence, and also, I think, in his Artificial Morality: How Morality is Rational (unpublished manuscript, draft 0.4, York University, 1988).

plans, commitments etc.? I argue that one can rationally choose and rationally act from dispositions, plans, etc., given one's preferences, but only where the actions which result from these things are maximizing. For the maximization principle, I shall argue, gives the structural form of all possible conclusive reasons for action, a structure violated by, e.g., Gauthier's proposals for constraining maximization. Thus it is not possible for choosing from a maximization constraining disposition or plan to comprise an alternative kind of rational choosing.

2. Gauthier

Gauthier aimed to resolve an apparent conflict between rationality and morality.[5] On the face of it, to be rational is to maximize one's individual Expected Utility (EU); to be moral, to refrain from maximizing. So how can it be rational to be moral?

He took this conflict to be modelled in the one-shot PD: each agent prefers, in descending order, unilateral confession, joint non-confession, joint confession, unilateral non-confession. If one agent, A, confesses, depending on what the other, B, does, A's best outcome is his first-best, his worst, his third; if A does not confess, his best outcome is his second, his worst, his fourth. If A's actions are causally independent of B's, no matter what B does, A's best and worst outcomes are better if he confesses; so if he is rational, he will confess. (This is the "dominance argument.") Likewise for B, so both normally confess. But then each gets only his third-best outcome. If only each would not confess, each would get his second-best outcome.

Here, the agents are conflicted because neither can get his best outcome without the other getting his worst. The non-moral solution is for each to seek his maximal individual advantage; the moral solution is for both jointly to seek their optimal mutual advantage. This involves mutual adherence to a moral rule or strategy for choice: do not confess. The defining property of moral rules: each agent does better if all comply with them than if all deviate, but better still if he deviates whether or not others comply. Call complying "Co-operating," deviating, "Defecting." The original problem of how it can be rational to be moral is now: How can it be rational to Co-operate when it is individually more advantageous to Defect?

Gauthier answers that while each agent has an incentive to Defect (confess), each, to avoid the disaster of mutual Defection, has a stronger incentive to arrange that all agents Co-operate, even if this secures his own Co-operation. But what of the temptation to Defect? Gauthier thinks agents can adopt dispositions constraining them from always performing individually maximizing actions. When facing a PD, it most advantages an agent to have whatever disposition will induce others to Co-operate, while yet allowing him to Defect as much as possible. Given perfect information, this is the one disposing

[5] This exposition of Gauthier is based on his "Morality and Advantage," The Philosophical Review, 76 (1967): 460-75; "Deterrence," "Afterthoughts," "In the Neighborhood," Morals By Agreement, especially Chaps. 1, 5, 6, and 11, and "Economic Man."

one to Co-operation just where, did it not, one would probably be Defected against, but since it does, provokes another to Co-operate, otherwise allowing one to Defect. Such agents do better than those disposed always to Defect, since they induce Co-operation from similar agents, ones who will Defect against those who will always Defect. They do better than those disposed always to Co-operate, since they benefit from the others' Co-operation, but further advantage themselves by Defection. Thus they reap all the advantages of both universal Defectors and universal Co-operators, incur none of their disadvantages, and enjoy further advantages available to neither. It is thus the maximizing and so rational disposition to adopt.

In choosing from it, one will Co-operate with a similar agent; otherwise, one will Defect. Agents with such dispositions are conditional Co-operators or "Constrained Maximizers" (CMers), conditional because they are induced to Co-operate only if the other agent is like disposed; constrained because their dispositions sometimes stop them from performing individually maximizing actions (from "straightforwardly maximizing," here, Defecting), and instead make them perform actions which, with those of other CMers, jointly optimize.

Gauthier claims then that in the PD, initially non-moral agents will find it rational to act morally. Each will see the advantage in all agents complying with choice strategies universal compliance with which optimizes; each will rationally commit to choosing from such strategies with others so committed, and will act on that commitment when facing them. For on Gauthier's conception of rationality, it is rational to act on rational commitments.

3. Problems With Gauthier's Theory

But what must CM dispositions (hereafter, CMDs) be like to make Co-operation on their basis free, rational, voluntary action, action chosen at the moment of the actual decision about whether to Co-operate or Defect, as Gauthier thinks it would be? In acquiring CMDs, either one's preferences remain as before, or they have changed. If the former, his account is incoherent. Think of the steps involved in making a final choice of action in Gauthier's PD: the agents first amend their characters, then assess each other's characters, then choose whether to Co-operate or Defect. Gauthier thinks standard maximizing rationality initially supports adopting the CMD before others assess one's character. For when other CMers see it, they will--supposedly--Co-operate, to one's advantage. (This is the only thing which makes it maximizing to have the disposition in the one-shot PDs we are considering here. There will be no future interactions, for instance, in which it would be an advantage to have proved oneself trustworthy in this interaction.) And Gauthier thinks that it is then rational to Co-operate from that disposition when later deciding whether to Co-operate, for that would express the rationally chosen disposition.

But surely the maximizing conception also justifies one in abandoning the CMD when one comes to choose between Co-operating and Defecting. For one's CMD has then already made others Co-operate (if it can). One now does best if one abandons it and then Defects. Thus the disposition rational (because maximizing) to have after one's character has been evaluated, and after one's character makes the other Co-operate, is one to Straightforwardly Maximize (the "SM" disposition). Rational agents will thus now acquire it, and Defect on its basis. (What changes is not the anticipated possible facts; both agents knew all the possible effects of their dispositions before, and know the same things after. But the consequences of having the CMD change: before the other chooses among actions, one's having the CMD can--supposedly--cause him to do something advantageous to one, namely, Co-operate; but after he has chosen among actions, the CMD can only have one further effect, namely, to make oneself perform the non-maximizing action of Co-operating. Thus the maximizing disposition to have after the other has chosen among actions would be one that would instead induce one to perform a maximizing action, not a constrained one.)

So even if rational action is chosen from a maximizing disposition, since, after one's character has been assessed, that disposition is the reverse of the one it maximized to have before, Defection is still rational. It would therefore be idle to adopt the CMD--it can make no rational difference to the final choice, and no one should Co-operate with anyone just because he happens to have it; for if he is rational and free, he will quit it. (We are assuming that the agents' bases for choice are transparent to each other, but that the agents' choices themselves are still independent of each other. The choice of action one agent will make is not conditional on the choice of action the other will make, but on the present and foreseeable basis the other agent has and will have for that choice. If their actions were dependent on each other, so that the second agent would Co-operate if the first did, it would, of course, be rational--because maximizing--for the first to Co-operate. But that would be a different game.)

Now, the CMD might be somehow conceived as permanent and irrevocable, thus really assuring others that one will Co-operate if they have it too. It would then maximize to adopt it, and it could then yield Co-operative behaviour.[6] But the rational principle justifying its adoption still justifies its later abandonment. Yet the disposition's stipulated irrevocability makes it impossible later to conform to that principle, which then insists on the disposition's abandonment. Thus, Co-operation from an irrevocable CMD is irrational (or perhaps non-rational, since one is not really controlling one's actions when it makes one Co-operate). One behaves Co-operatively only because one is made to by the disposition, not because it is rational to have it and to act on it when it forces one's hand; it was rational to have it before one's character was evaluated, but not after. Thus the principle instructing one to act on dispositions which it is maximizing to have is incoherent, for it insists on adoption of an irrevocable CMD (else it cannot assure reciprocal Co-operation), and also on its later (expressly impossible) revocation, when it would maximize to adopt a Defector's disposition. (Would it be better to say that the CMD is still the rational disposition to have, given that one had to choose an irrevocable one to maximize in the initial choice among dispositions, and given that one can hardly

[6] This repair is suggested in Richmond Campbell, "Moral Justification and Freedom," The Journal of Philosophy, 85 (1988): 192-213, criticized in detail in my "Libertarian Agency."

be called irrational for not doing the impossible? By definition, one cannot change an *irrevocable* disposition. So how could rationality oblige one to do so? Good question. But my point is that on the current formulation, the impossible is just what Gauthier's principle requires one to do. That is my objection.)

If dispositions leave preferences intact, yet really give others reason to think one will Co-operate, since that involves one going against one's continuing preferences (even, it seems, on Gauthier's conception of what it is to "go with" one's preferences), one's dispositions must causally force one to Co-operate. They thus merely cause, not rationalize, Co-operation.

So if it is always rational to act from the disposition which it maximizes to adopt, then one should Defect, since after the other has acted, that disposition becomes the disposition to Defect. Since both agents would both know this, if both remain free to choose, neither will Co-operate. So both would still do poorly (if they could always choose freely!), even in Gauthier's rationality.

Gauthier might reply that were one going to revise one's disposition after the other Co-operates, one could not make him Co-operate by initially disposing oneself to conditional Co-operation. So one should adopt a disposition which will not be revoked, then Co-operate from it. Both agents will then Co-operate, doing well. And he might claim that one acts freely in Co-operating, since the principles of rationality make the Co-operative choice the natural one. Co-operation from a rational commitment to Co-operate is what a rational agent would choose even given the choice to do otherwise. This suggests another interpretation of Gauthier's proposal: *(c) a choice is rational if concordant with the disposition it maximized to adopt for the choice*. If it was rational to commit to doing *x*, it is rational, in virtue of the commitment, to do *x* when the time comes.

But the same objection applies. Before having my character assessed, I dispose myself to Co-operate. Rationality apparently will license me to choose from that disposition after the other chooses his action. Suppose he has chosen. I now reflect on what it would be rational to do. I see that the disposition which it would now maximize to adopt for the choice among actions is one to Defect, and that *(c)* justifies me in acting from *that* disposition for that choice. So I must Defect. Or if not, at best, *(c)* gives me contradictory advice on how to choose: it justifies me in Co-operating from the Co-operative disposition which I earlier adopted, and in Defecting from the Defector's disposition which it would later be rational for me to adopt. What to do? Surely since the second is rationality's most recent deliverance, the one updated to reflect my new circumstances (in which my disposition can no longer affect the other's choice), I should Defect. The only reason I might not is that perhaps I cannot (since the CMD had to be irrevocable to maximize at the time of adoption); but then I am not doing the (currently) rational thing in Co-operating. If the CMD is irrevocable, both agents will Co-operate, doing well, but they will not be acting rationally.

Now it may seem that Gauthier wants us to be backward-looking in deciding the rationality of a current choice, and that I have been insensitive to this desidiratum. If one made a maximizing commitment earlier to choose a certain way later, one should regard that as decisive in deciding how to choose later. Maybe so. But my point is that if what makes a current choice rational is that it follows from a commitment which it maximized to make before the choice, there is another such commitment which it is maximizing to make after the first commitment, but still before the final choice among actions. If one is rationally obliged to choose on the basis of a commitment which it earlier maximized to make for the choice, surely either there are too many pre-commitments from which it is supposedly rational to choose, or there is a principle for determining which commitment it is rational to regard as decisive. And I claim that, since what is supposed to make a commitment rational is that it maximized to make it, and since the second one is the one that most recently maximized, that is the one that one should follow (if one could; if one had a choice).

If (c) gave a unique verdict on how to act (e.g., if it said only that it is rational to act on the irrevocable disposition which it was rational to adopt before the other agent chose among actions), then Co-operating from the CMD might be a free and rational choice. We could just redefine such a choice so that it is not one that maximizes on current preferences, but one that expresses the disposition which it earlier maximized to adopt. But (c) does not do this; rather, it also requires one to revise one's disposition (whether or not one can) after the other has acted.

Gauthier might reply, however, that were it rational to revise one's disposition after the other Co-operated, and if one could always act rationally, then since the other would foresee one's revision, he would not Co-operate. But true rationality would allow one to make the choices maximizing overall. They are choices first of a CMD, and second, of Co-operation from it even after the other has Co-operated, since the first choice will induce the other to Co-operate, given that it will induce the first agent to do so, for a better result than from one making choices that will induce the other to Defect. Here, both agents will Co-operate, doing well, and both will, Gauthier would think, be rational in doing so. This suggests his proposal is really: (d) a choice is rational if it expresses the disposition the having of which maximizes overall. One asks which would be better, to choose dispositions that would permit a later Defection, or ones that entail Co-operating. Truly rational agents could always do what would most advantage themselves, and committing to Co-operation is more advantageous than making choices that will later entail Defecting.

But while the disposition maximizing overall (i.e., from now on) before the other decides whether to Co-operate is a CMD that will not be revoked, the one maximizing overall after (i.e., from then on), is the disposition to Defect. And rationality, surely, is continuingly responsive to relevant changes in circumstances. What makes a disposition rational is surely that having it maximizes overall in present and forseeable circumstances. But these have changed after the other Co-operates (since that is now in the past of any of one's choices and of any of the future effects of one's dispositions, whereas before, it was in their future), thus so has the disposition maximizing overall. So one's disposition rationally must change even by (d). Both agents, knowing this, will Defect if they can, doing poorly. If one cannot change one's disposition because the original one is irrevocable, as it must be to be advantageous, this does not mean that

Co-operation from it is unequivocally rational, for by (d), there is another disposition which rationality would oblige one to adopt after one has adopted the irrevocable one, and after the other agent has acted. Option (d), then, is as equivocal as (c).

One more try. Perhaps what Gauthier meant was that whenever it is necessary to procuring a certain advantage that one commit to a certain action, e.g., to choosing in such a way as to forego a certain future possible other advantage, one must, rationally, choose by that commitment. Thus: (e) a choice is rational if dictated by the disposition/commitment which it maximized to undertake if that undertaking was a condition of enjoying an earlier advantage.

Sadly, however, our objection returns again. It initially maximized to commit to Co-operation. One might then think one is rationally obliged by (e) to Co-operate. But after the other chooses among actions, there is an advantage to be had by committing to Defection, namely, guaranteeing the expected utility of Defecting whatever the other has chosen. Since disposing oneself to Defect after the other chose among actions was a condition of enjoying the advantage of expecting the utility of Defection, it is rationally obligatory by (e) to Defect (or one has, again, a dilemma about which commitment it is rationally obligatory to honour); and around the loop we go again.

I suggest that no version of Gauthier's proposal will rationalize Co-operation. For while he tries to commit agents to Co-operative choices by calling choices from rational commitments rational even if the choices are not themselves maximizing, the commitments which it is rational (because maximizing) to make vary with the circumstances. Gauthier still decides the rationality of a commitment by whether it maximizes, and the fact that the maximizing commitment changes in the circumstances for which earlier commitments were to determine actions, means that the old ones cannot always rationally determine actions for the times to which they were designed to apply.

4. A New Solution: Co-operation Rationalized

I noted earlier that in acquiring a CMD, either one's preferences remain as before, or they change. The former had problems. But consider the latter: in acquiring a disposition (in committing) to Co-operate, one acquires a revised set of preferences, given which it straightforwardly maximizes to Co-operate.[7] I suggest that one should choose the

[7] I advocate this in my "Two Gauthiers?"--though promisorily and in much less detail than I will here. The idea that it is rational for agents to revise unilaterally their preferences should not be confused with Amartya Sen's proposals in his "Choice, Orderings and Morality," in Practical Reasoning, edited by Stephan Korner (Oxford: Basil Blackwell, 1974), p. 54-67, nor with those in Sen's "Reply to Comments," ibid., p. 78-82, and "Rationality and Morality: A Reply," Erkenntnis, 11 (1977): 225-232; nor with those of Edward F. McClennen in his "Prisoner's Dilemma and Resolute Choice," in Paradoxes of Rationality and Cooperation; Prisoner's Dilemma and Newcomb's Problem, edited by Richmond Campbell and Lanning Sowden (Vancouver: University of British Columbia Press 1985), p. 94-104. For a summary review of the differences between their proposals and mine, see my "Two Gauthiers?", p. 54-55. For a more

preferences which it would maximize to have, given one's original preferences, then choose actions from those new preferences. One will choose preferences to Co-operate with those with like preferences, then choose Co-operation from the preference to Co-operate with them. This has the same effects as Gauthier wanted his "dispositions" to have, and it is undertaken for the same reason (to provoke others to Co-operate), but it rationalizes the resulting choices by the classical standard that one should always maximize on whichever preferences one has when choosing. This recommends choosing preferences for Co-operation before the other chooses among actions, and choosing Co-operation after. Both agents will Co-operate, doing well by their original and their later preferences; both will be rational both in revising their preferences and in acting from the revised ones. We can now unpack--indeed, demystify--being disposed to Co-operate: it is being in a psychological state composed of preferences and beliefs. If you like, this rehabilitates Gauthier's proposal, but requires him to withdraw his claim that Co-operation issues from constraint. It is really just more maximizing behaviour. Thus my conception of rationality: (f) a choice is rational if it maximizes on the rational preferences one has when choosing; preferences P are rational if having them maximizes their own satisfaction, i.e., if there are no other preferences, P*, such that having them would better advance satisfaction of P.

We shall explore the character of this proposal in a moment, but first we must pause to consider an important possible objection.

5. Maximization and Morality

Objection: Showing that it is rational to change one's preferences so that it maximizes to non-confess in a PD does not prove the rationality of morality. For moral behaviour is, *per* Gauthier, *constrained* behaviour. *Were* it maximizing to non-confess, non-confession would not be *moral*, only *prudent*. Indeed, I do not even show that it is rational to *Co-operate*, for *per* Gauthier, Co-operation must be *non-maximizing*, while I try to make it maximizing.[8]

Reply: Both objections are misplaced. Gauthier sought to capture two features of morality. First, moral actions are ones which one ought to perform whether one initially wants to or not.[9] Second, moral actions implement strategies for optimizing the welfare

detailed critique of these early attempts to rationalize Co-operation in a PD, see my "Co-operative Solutions to the Prisoner's Dilemma," *Philosophical Studies*, 64 (1991): 309-321. My proposal *is* similar in certain ways to one in Edward F. McClennen, "Constrained Maximization and Resolute Choice," *Social Philosophy & Policy*, 5 (1988): 95-118; and what I am saying here can be read as an elaboration and defence of his view as well.

[8] My thanks to an anonymous referee for these objections.

[9] I repeat: whether one *initially* wants to or not. Gauthier himself sees agents coming to prefer moral conduct, ceasing to experience it as constrained. But for him--unlike for

of agents in resolving partial conflicts of interest.

My account captures the second feature, for choosing as I recommend will result in at least optimality. My account also captures the first feature, but to see how, we must say more about Gauthier's conception of morality. In effect he analyzes moral behaviour as implementing a strategy of choice which, if both agents do their part in it, optimizes their utilities, and which resolves their partial conflict of interest. One can calculate what it is moral to do by what it jointly optimizes for both agents to do. Agents adopt moral strategies for self-interested reasons, and then act on them because they have adopted them. Now, being moral is doing the right thing for the right reason. One might think (though I do not) that the right reason for the right action cannot be that the action maximizes. But Gauthier's agents only adopt "moral" dispositions because that maximizes. Thus does not action from his "moral" dispositions ultimately have maximization as the motive? No. His agents only have maximizing reasons for adopting their dispositions, not for acting from them. For it is not maximizing to keep to an agreed constraint. Rather, Gauthier analyzes as what makes an action moral that it is dictated by a rationally agreed-upon constraint, and then insists that the only agents who can benefit from such constraints are ones who can have as a reason for an action that it is dictated by an agreed-upon constraint. Since agreed-upon constraints represent what is right, his agents thus do what is right because it is right, not because it maximizes. So his agents do moral actions for moral reasons (even--indeed, especially--on the account of right reasons as reasons unrelated to maximization). For Gauthier, what makes actions just is that they are dictated by maximizing dispositions, i.e., ones which all rational and relevant parties would find rationally agreeable. So dispositions are just because maximizing; actions are just because they follow from just dispositions. Thus, for Gauthier, just actions are good because they are just, not because they maximize (though maybe because they socially optimize, but that is different, and prima facie moral). The result: Moral acts are ones which we rationally ought to do whether or not initially we prefer to.

My account also has it that one ought to do certain things whether one initially wants to or not, namely, those conforming to rationally agreed-upon optimizing strategies in partial conflicts. Agents acquire a preference to do their part in an optimizing joint strategy because so preferring is in their original interests. They act optimizingly because they have come to want to. Their reason for non-confessing is that it is moral (i.e., mutually rationally agreeable) and they now prefer to do what is moral. But while Gauthier tries to show that agents ought to do moral actions by rationalizing adoption of, and action upon, dispositions constraining agents from maximizing, I try to show it by rationalizing adoption of, and action upon, new preferences, ones which make such actions maximizing.

me--one does not prefer it as a condition of its rationality, but merely as a means of reinforcing moral conduct already made rational by one's having rationally adopted a disposition to be moral. For details on the differences between our proposals, and on whether he himself can be interpreted as offering the preference-revision solution I favour, see my "Two Gauthiers?".

But how can maximizing actions be moral? Well surely it would be odd to so define moral actions as to make them ones which agents do not want to perform, ones that must not advance their interests no matter how noble or other-regarding those interests. Even Kant did not require that one's moral actions not be in one's interest, only that the motive or principle of action not be individual interest; an action is only moral if performed because it is moral, regardless of whether it is in one's interest. Now if the interest of the PD agent is, initially, to minimize his jail time, then when he refrains from confessing on my rationale, he does not act in that interest. For it is not refraining from confessing that serves it--better by that measure to confess. So when he refrains, he chooses against his original interests. What serves those is his adopting preferences which make non-confession maximizing in interactions with certain types of agent (because his so adopting makes those agents Co-operate with him). Of course he now has new interests, which is why he now refrains. Still, he does an action not in his original interest; and when he acts, he does it because it is right (i.e., mutually agreed-upon). But since he now prefers to do what is right, doing it happens now to be in his interest.

To think that moral conduct must not express an agent's preferences would be to commit the sophomoric fallacy of assuming that actions expressing one's preferences are necessarily selfish actions, not actions that have as their aim respecting moral obligations or enhancing the welfare of others. But what distinguishes selfish actions from moral actions is not that moral ones do not express the agent's preferences, only that they express preferences which cannot be satisfied except by compliance with a moral principle, or by enhancements to others' welfare. The preference to Co-operate meets this standard; it can only be satisfied by compliance with a moral principle, compliance with which enhances another's welfare. Notice, then, a crucial difference between the motives which agents have for moral conduct on my account contrasted, say, with Hobbes's. Hobbes's agents cannot internalize a motive of action that has as its objective the welfare of others or conformity to a moral principle. They only act morally from fear of a penalty to their own welfare for non-compliance. My agents, by contrast, have internalized a motive aiming at others' welfare or at moral conduct. They have come to want to do the right thing (Co-operate) because it is the right thing (the thing all parties find rationally agreeable). Put another way, where Hobbes's agents only do the right thing for selfish reasons, my agents do it because they have come to care about doing the right thing for its own sake. His agents cannot do the right thing for what we would call the right reason; mine can.

Finally, on the nature of Co-operation: my way does rationalize Co-operating proper, since what defines Co-operation is not that it is non-maximizing, but that it implements a strategy optimizing on the initial preferences of PD agents in partial conflict. But what of Gauthier's claim that it maximizes to commit to moral strategies of choice, but not to act on them? Again, I preserve this: it maximizes on one's original preferences to revise them--to come to prefer to act on the optimizing strategy--but it does not maximize on the originals to conform to the optimizing strategy--to maximize on the new preferences to act on that strategy. In so complying, one only maximizes on the new preferences, but since one no longer has the old, one does not mind that they are not advanced. But while one's new actions do not maximize on one's old preferences, one never constrains oneself from maximizing on any of one's preferences. Rather, one maximizes on the originals in revising them, and on the new ones in non-confessing.

Surely philosophers only ever wanted a demonstration of the rationality of constraint because, since Hume, they could see no rational basis for criticizing preferences: there is nothing irrational about preferring to avoid the scratch of a finger over saving humanity from a holocaust. Thus they thought one could never show the rational preferability of morality; for (short of coercion or socialization) one could never give agents with non-moral preferences, preferences to act morally. Thus philosophers were driven to try to give agents reasons to restrict themselves from always pursuing their own interests, from advancing their preferences. (Indeed, making virtue of necessity, perhaps the doctrine that agents might not find it preferable to act morally got reified into a dogma, into an analysis of what it *is* to act morally: moral behaviour is *essentially* non-preferred behaviour, behaviour that expresses a constraint on the expression of individual preferences.) But I claim that instrumental rationality entails the rationality of changing one's preferences in PDs. It is in agents' interests to prefer not to exploit (certain) others. Thus, in certain contexts, there *is* a standard of criticism for the rationality of preferences; and PD agents initially have the *wrong* preferences by that standard. Their initial non-moral preferences in the circumstances make it rational for them to adopt moral preferences.

Let us continue then with our exploration of this proposal in detail.

6. The New Solution Applied in Detail to the PD

When about to face a PD one should adopt a new preference-function, knowing other agents will choose in light of it. But *which* function? We can make progress on this question by conceiving the situation as a non-standard PD, one that is really a two-step game. In the standard PD, the agents choose among Co-operation and Defection, given their beliefs and preferences. In this PD, they first choose among preference-functions, given their beliefs and preferences, then choose among Co-operation and Defection, given their beliefs and new preferences. (Gauthier's is also a two-step PD. But in the first step the agents choose among dispositions.) To choose rationally in the first step, each agent must choose preference-functions, the having of which maximizes by the standards of his initial preference-functions. That is, he must choose preference-functions which will induce other agents to choose in ways most favourable to him by the standards of his initial preferences, while allowing him to choose favourably to himself as often as possible, given this. (To choose rationally in the second step, he must maximize on the preferences chosen in the first step.) To give yourself the highest utility by your original preferences, you want a preference-function which will make others Co-operate, yet let you Defect. But no rational agent with PD values will knowingly adopt preferences which would make him Co-operate with agents whose preferences would then make them Defect. So the best you can have is a preference-function which will make others Co-operate with you, and you with them exactly when, if you would not be made to

Co-operate, they would not be either, one which will make you Co-operate just where its doing so makes them reciprocate, otherwise making you Defect. (By the way, when I speak of agents' preferences "making" them choose in certain ways here, I precisely do not mean "making them choose a certain way against their will." I mean that their preferences rationalize their choosing a certain way so that, so far as they are rational, they would choose that way. In so choosing they express their will; they do what they want--just the reverse of what happens when one's behaviour merely conforms to the dictates of a constraining disposition.)

An agent about to enter a PD normally has the following preferences in the following orders (C=Co-operate, D=Defect, his action, the left-most in each pair, his partner's, the right. Each letter is an action, each pair, an outcome, the order of pairs from left to right, an ordering of outcomes from most to least preferred): DC, CC, DD, CD. One might interpret my proposal as arguing that each agent rationally should revise his ordering to look like this: CC, DC, DD, CD.[10] And one might think that when such agents meet they will each Co-operate, since each most prefers the outcome of joint Co-operation.[11] But this will not always work. For suppose each notes that each prefers outcomes in which both have Co-operated. Has either a reason to Co-operate? Not if their choices will be made in secret. There is no dominance argument for Co-operating, for instance, nor necessarily a maximizer argument. The agents are you and me. I get 40 utiles for my most preferred outcome, 30 for the second, 20 for the third, 10 for the fourth; same for you. If I Co-operate, assuming a 50% chance that you will Co-operate, my EU is 50% of the utility of CC (20) plus 50% of that of CD (5), total, 25. If I Defect, my EU is 50% of DD (10), plus 50% of DC (15), total, 25. Since the EUs of Co-operating and Defecting are equal, it is indeterminate how I should choose. (If I think it more likely you will Co-operate, the EU of my Co-operating rises, and I should Co-operate; if less likely, Defect. But I have no reason to assume either. It is not that I lack the information needed to predict your choice, given your basis for choice--given your preferences, your beliefs and your rational principles of choice, given both. It is that nothing in your beliefs, preferences or principles makes it uniquely rational for you to choose one way or the other.)[12]

Things are better if our choices among actions are public. For if you have Co-operated and I know it, I have reason to Co-operate, for that gets me the joint Co-operative outcome I most prefer. Likewise for you if I have Co-operated, etc. If

[10] These preferences define Amartya Sen's Assurance Game. See Sen, "Choice."

[11] Richmond Campbell thinks some of my earlier work implies this proposal and has his own objections and alternatives to it. I am not sure my earlier work is clear enough to imply it; but in any case, I think it is probably wrong--see below, in the main text.

[12] For a detailed critique of these and various other preference-functions that some philosophers have thought would rationalize PD Co-operation, see my, "Co-operative Solutions."

neither of us has chosen, given the opportunity to choose sequentially, I know that I can provoke you to Co-operate by Co-operating myself; same for you. So we will both Co-operate.[13] But to deal with our secretly choosing actions after publicly choosing preferences, we must each publicly adopt the following preferences with the following ranking from most to least preferred. (We assume the agents will know each others' preferences and that each other are rational, but that their choices among actions may or may not be secret.) Each agent should prefer to: (1) Defect against (i) anyone who he knows did, will, or likely will Defect,[14] (ii) anyone unconditionally disposed to Co-operate, and (iii) anyone unconditionally disposed to Defect; (2) have outcome CC; (3) Co-operate with just (iv) those disposed to choose as if their first two preferences were, in order, (1) and (2) and who do not fit (i); (4) have outcome DC with agents who satisfy (iv); (5) have outcome CD. He will then Defect on those who fit (i), (ii) or (iii) for that directly maximizes given his strongest preference, (1). But even if actions are secret, with someone, B, who fits (iv) and not (i), he thinks: "I cannot Defect on the rationale of (1) because B does not fit (i), (ii) or (iii). I can satisfy (3) by Co-operating because B fits (iv) and not (i), while if I Defect, I can only satisfy (4). So I have sufficient reason to Co-operate. But I have even more reason in that my doing so will likely help satisfy (2), since B also has those reasons to Co-operate. So, *a fortiori*, I should Co-operate."[15]

Unlike the original preference-function, this one does not contain only preferences for outcomes *qua* pairs of *choices*. It discriminates more finely among the possible worlds which are the objects of preferences. It does this on the basis of the preferences or dispositions of the other agent and the actions which each agent has performed in each possible world (or outcome). Thus our agent prefers first, worlds where he Defects against someone known to be disposed always to Co-operate, or disposed always to Defect, or who it is known did, will, or will likely Defect; second, worlds where both agents have Co-operated; third, where he has Co-operated with agents disposed to choose as if their first and second preferences were like his first and second preferences; fourth, where he has Defected, the other has Co-operated, and the other is disposed to choose as if his first two preferences were like the first agent's first two preferences; fifth, where he has Co-operated and the other has Defected.[16]

[13] For more on this, see my "Co-operative Solutions."

[14] We need this clause to cover the case where one discovers just before one chooses actions, after having chosen one's preferences, that the other agent did not, will not, or will likely not, act on his preferences; this clause gives one a basis for protecting oneself by Defection, should one learn that the other was not perfectly rational, or was prevented from acting rationally, and so failed to act as he should have, given his preferences.

[15] Some other aspects and defences of this solution are given in my "Co-operative Solutions."

[16] My thanks to Terry Tomkow for help in formulating the proposal in these terms.

The agent does not come to be inclined to choose in a different way, given his original preference-functions over outcomes; he still chooses as a maximizer. But his preference-functions have changed. He used to prefer unilateral Defection no matter what. But he now only prefers it against suckers, cheaters and accidental Defectors. These new preferences are not meta-orderings, if by that is meant orderings on orderings.[17] It is not as if one prefers outcomes in a certain order, then prefers to act as if one preferred those outcomes in a different order; I think that that would result in an incoherent overall preference ordering.[18] Nor are these preferences over path-dependent outcomes, preferences in which it does not just matter to one what happens, but also how it happens. What matters to one is outcomes, but part of those outcomes is what action one performed, and the circumstances in which one performed it--circumstances defined, in part, by the actions and inclinations of others.[19]

Here, then, is a preference-function, adoption of which advantages you by your original standards in a one-shot PD, given the transparency of agents' characters, with a pre-interactive opportunity for their amendment, assuming your basis for choice can influence the other's choice. Of course, you would do better still by your original values (if you still had them) by Defecting against even those who will Co-operate with you if you prefer to Co-operate with them. But since your first relevant preference is now to Co-operate with them, you will not Defect. You knew that so changing your preferences deprives you of Defection as a rational option in that case, but that is a good bargain in your initial overall calculations; preserving the option of Defecting against a conditional Co-operator (someone who wants to Co-operate with you provided you have certain preferences) would doom you to your originally third-best outcome, so you happily forego it for a better chance at your second.

More generally, whenever we cannot both get our best outcome, and can only get our second-best if we co-ordinate, we should come to prefer performing our half of co-ordination to achieving the originally best outcome, with those who prefer that we both have co-ordinated over achieving their originally best outcome. We should come to prize co-ordination with certain types of agents over the original best outcome, and to prefer whatever would most conduce to the originally first-best outcome whatever the other chooses if he does not prize joint co-ordination sufficiently highly--or if, for some reason, it is expected that he will not (or suspected that he did not) act on that value.

[17] Thanks to Christopher Morris for asking.

[18] See my "Co-operative Solutions."

[19] Again, thanks to Christopher Morris for asking.

7. Freedom From Circularity

This solution avoids three kinds of circularity that can damn solutions to the PD.[20] First, each agent's preference is specified non-dependently on the other's, avoiding the problem of neither's being determinately defined. We have symmetry, but no circular dependence of definition. (We would have the latter if we defined the preferences this way: "We each prefer to Co-operate, provided the other prefers to Co-operate." But then neither of us yet prefers to Co-operate.) Each prefers the mutually Co-operative outcome to all but Defection against habitual suckers and cheaters (and those who accidentally Defect), and prefers to Co-operate except with those inclined always to Co-operate or always to Defect (or who accidentally Defect).[21]

Second, each has reason to acquire the preference-function independently of whether the other acquires it, for it does not disadvantage him against the unrevised, it allows shared advantage with the appropriately revised, and it allows him to exploit chronic Co-operators and accidental Defectors. So each is not left waiting for the other to revise.

Finally, each may rationally choose an action independently of the one which the other will choose (assuming the former agent will not know of the latter's choice before the former makes his choice), so they are not left waiting for each other to choose. Each does not act on his new preferences only if the other does too, but just if both he (agent A) and the other (B) are mutually known to have the appropriate preference-functions and to be rational and free, and so likely to act on their preferences. Of course, A predicts B will Co-operate, and acquires the new preference-function only in the hope that, should A meet a B with a similar function, A's function will stimulate B to act on B's similar function. But when A Co-operates, he does so not because B will, but because B has a similar preference-function, and A has come to want to Co-operate with anyone like-functioned.

[20] My thanks to Peter Danielson, who worried in his referee's report for "Two Gauthiers?" that my CM preferences "violate conditions of independence required of standard preferences, since a CM preference for Co-operation is dependent on the other's similar preference."

[21] For a defence of other, Gauthier-type solutions from the charge of circularity (but ones which leave preference-functions unchanged), see Richmond Campbell, "Critical Study: Gauthier's Theory of Morals by Agreement," The Philosophical Quarterly, 38 (1988): 343-364, and Peter Danielson, "The Visible Hand of Morality," (Review of Gauthier, Morals By Agreement), Canadian Journal of Philosophy, 18 (1988): 378-379.

8. The Determinacy of the Solution

But if one is revising one's preferences, why not just adopt ones satisfiable whatever the other does (e.g., a preference always to Co-operate), and which, when satisfied, would yield huge individual utility? Why not come to regard one's life as having achieved its apogee if only one Co-operated (or Defected, for that matter)? But then, is not the correct solution to the PD underdetermined by rationality? If a way out of a PD is simply to revise one's preferences so as to achieve maximal utility, since there is no limit on utility in principle, the choice of preferences could only be fixed by something other than rationality, like what one's contingent psychology limits as one's maximum possible happiness or restricts as the sorts of preferences one can acquire. And is this not a reductio of the very idea of there being a determinate rational solution to the PD?

No. This misconceives the relation between rational choice and utility. To say the rational agent is a utility maximizer is not to say he wants to give himself as much happiness as possible, as if happiness qua utility were something he could prefer, something he would, rationally, have to give himself as much of as possible by whatever means. It is elliptical for saying he always chooses so as to maximize the product of the likelihood and the cardinally ranked desirability of the outcome of his choice. (If we have only ordinal rankings, agents seek to maximize the product of the likelihood and one over the ordinally ranked desirability of outcomes.) Thus the utilities attaching to outcomes do not comprise some intrinsically preferable commodity; they just specify what one prefers most, second-most, etc., and in what strengths. PD agents initially want nothing but as little jail time as possible; or, put in terms of ordered preferences for states of affairs in possible outcomes of choice, first, to have Defected where the other Co-operates; second, to have Co-operated where he Co-operates, etc. Thus they are only justified in making such revisions to their preferences as would most likely cause those outcomes in the order originally preferred. Preference-revision is justified, assuming that one's preference-function can influence another agent's behaviour, make him help bring about one's most preferred outcomes. Coming to prefer always to Co-operate or always to Defect would not do that.[22]

[22] My thanks to Peter Danielson, Robert Bright, Terry Tomkow, and Sheldon Wein for this worry, which plagued an earlier version of this proposal. I hope this also begins to meet concerns tentatively expressed in Danielson, Artificial Morality. Danielson worries that permitting the revision of preferences makes the conditions of choice unstable; also that players will find themselves in a vicious regress of meta-games, each inflating his preference to drive up the price of his concession on something also valued by the other agent. I do not think our PD creates this problem, for, as I argue below, the agents have an interest in co-ordinating their preferences, not in trumping those of the other agent.

9. Change of Preferences or Principles?

Principles are functions from preferences to choices of action, given beliefs; preferences are attitudes to outcomes, or functions from outcomes to degrees of satisfaction. I say the correct solution involves revising one's preferences. But for several reasons, one might think it really involves principle revision. First, I say the rational agent would both prefer to Co-operate with those who prefer joint Co-operation, and for his individual jail time to be minimal. But you cannot rationally prefer both to Co-operate instead of minimizing jail time, and to minimize jail time instead of Co-operating. So unless I am recommending ill-ordered preference sets, the "preference" to Co-operate must be construed as a principle of choice, something that does not compete with the preference for minimum jail time as a determinant of choice, but which tells one how to choose, given it.[23]

But we escape this easily enough: the preference to Co-operate just with those with a preference for joint Co-operation must be ranked above the one for minimal jail time. Now the former can be a preference, and can rationally coexist with one for a minimum jail time, for it is merely a stronger one; one prefers Co-operation with those with a certain preference-function to minimal jail time, and minimal jail time to everything but Co-operation with those with the right preference-function.

But this raises a second worry. If one so-called preference for action really overrides the other (a preference for an outcome), the first has the character of a principle; it figures not as an attitude towards outcomes, but as a way of choosing, given such preferences.[24] So, on the above definitions, an attitude towards an action or a choice is not a preference. I think, however, that we may still consider it a preference for an outcome merely by counting which actions have been performed as part of the outcome along with the resulting jail sentence. One prefers to have Co-operated in the outcome with those who prefer joint Co-operation, more than one prefers to have Defected and received an even lower jail time. Moreover, in a rational psychology, preferences only come ordered; so it is no argument for something's being only a principle that one ranks conformity to it above satisfaction of some preference. This just means one prefers what it is a preference for over what the other is a preference for.

True, the preference for actions which I endorse has one property of a principle: it is a function, given beliefs, from preferences for outcomes to choice of action, since if one

[23] Thanks to Peter Danielson and Richmond Campbell for these worries, also found in the notes to Lewis, "Devil's Bargains."

[24] There is something odd about this conclusion, for if the second so-called preference is really a principle, it is not sensitive to variations in the content of the first preference in what actions it recommends. Rather, it kicks in whenever the first cannot be satisfied, whatever it is. Compare with the maximizing principle, which gives different advice for action depending on the preference's content. But never mind.

prefers to have done something, ceteris paribus, one prefers to do and will do it (if rational and free to do what one prefers to do, i.e., what it would be maximizing to do). But there is no reason why preferences cannot determine actions. And what I have in mind is a preference rather than a mere principle because, as I have it, its satisfaction (acting as it requires) itself generates utility for the agent (as it were), for he now prefers such action for its own sake (or better, we can now represent his relative preference for it by assigning it a high individual utility value).

It might be objected that the "preferences" which I recommend are too conditional to merit the name. Normally, I would rationally come (on instrumental grounds via the dominance argument) to prefer to Defect, given my preference for minimum jail time, no matter what your preferences. But here, it seems I prefer to Co-operate if you prefer a joint Co-operative outcome, to Defect if you do not. This misrepresents my view slightly (see next para graph), but in any case, I think all preferences are conditional in this way. Take, for example, the original PD preferences: if you prefer non-Co-operative outcomes, I come, on instrumental grounds, to prefer to Defect, for that makes my best and worst possible outcomes as good as possible; if you prefer Co-operative outcomes, again, I prefer to Defect, for the same reason. All that changes in (this construal of) my proposal is the second clause in the foregoing conditionals: if you prefer the joint Co-operative outcome, I prefer to Co-operate.

However it might be claimed that on my proposal, one does not immediately adopt a new preference, but merely a recipe or principle for what to prefer, given what the other prefers. Thus one becomes such that: if he prefers a joint Co-operative outcome, one prefers to Co-operate, elsewise, not. But what I actually have in mind is the following. One comes to prefer that: if he prefers a joint Co-operative outcome, one Co-operates, and if he does not, one Defects. Note the difference in the scope of the preference. One's new preferences are fully formed at the moment of revision in the basis of one's further choices; thereafter, maximization on them selects specific actions, depending on the other agent's preference-functions.

10. Change of Principles as an Alternative Solution?

But even if principle revision is not my solution, could not a rational agent revise his principles, keeping his preferences fixed? Could he not continue to prefer minimal jail time, but acquire a principle like, "jointly optimize utility with others whose principle is to so optimize, otherwise, individually maximize"? This may be what Gauthier had in mind all along (with his talk of a Co-operative or joint strategy). If one acts on one's old preferences, given one's new principles, one will Co-operate with similar agents, getting one's second-best outcome by one's continuing preferences.

But why act on one's principles if doing so goes against one's preferences? It seems that rationally modifying one's principles is just becoming irrational (and while it may be rational to choose to become irrational, one's later choices will be irrational; but we are trying to see how they could be rational). For consider how rational agents justify their actions. Given well-ordered preferences, an action is rationally justified just if, first, one prefers to perform (or to have performed) it for its own sake sufficiently strongly as to prefer to perform it all things considered; or if, second, one thinks it advances satisfaction of a preference for something which one prefers for its own sake, all things considered. In the former, one so acts because one prefers to, simply for the sake of so acting; in the latter, because one thinks that will get one what one prefers for its own sake. One cannot justify a choice by showing it consisted in what one least preferred to do for its own sake, or least advanced getting what one most preferred, all in. That would be paradigmatic irrationality. Yet irrational is just what someone would be if he Co-operated on the putative rationale of his new principle, while still most preferring the outcome of unilateral Defection.

What if he Co-operated because that followed from an earlier choice among principles which he made for standardly acceptable maximizing reasons? Then we would ask why he did not change his principle when the expected utility of having it changed. (Remember, a principle is here justified just if having it maximizes.) What if he did not change it because at that point he could not? Then we would say that he was helpless not to act irrationally (or non-rationally), though he may have had good reason to bind himself.

To be rational, one must do what one prefers, or what one thinks will get one what one prefers. And this is captured in the maximizing conception of rational choice, except sensibly adjusted for the ratio of degree and chance of cost to degree and chance of benefit. This precludes from the outset functions from preferences to actions alternative to the maximizing function as the defining principle of instrumental rationality. And this goes very deep. People, simply as rational agents, do not thereby necessarily aim at conforming to principles (unless there is a principle they happen to have a preference to conform to); rather, their principles characterize how they choose, given their beliefs and what their preferences fix as what they aim at getting. An agent is not made rational by preferring to conform to rational principles, but by his actions and intentions being consistent with a principle for determining his choices, given his preferences (and beliefs). (This much, at least, accords with Gauthier's idea of instrumental rationality.) The principle is: maximize expected satisfaction of one's preferences. The deep part (with which Gauthier seems not to agree) is that the maximizing conception gives the very idea of having a conclusive reason for action. Conformity to it is what makes action from a conclusive reason different from behaviour from a mere cause. For it is the only principle on which preferences justify behaviours, in which the phrases, "I did it because I wanted to do it," and, "I did it because I had to do it to get what I wanted" come out true of one's actions.

I am often accused of begging the question against Gauthier when I say that rational agents do what they most prefer, implying that at best, agents who act as Gauthier intends from his proposed basis for choice would be guilty of rationally motivated irrationality. However I do not think I beg the question against Gauthier in my internal criticism of his views; I simply try to show that by his own stated measure, Co-operation is irrational. And this is because it turns out that the measure has a consequence he never intended: a choice is rational only if it or its consequences are preferred. This becomes an obvious consequence of his view that a choice is rational if it expresses a maximizing disposition, once we see that when a disposition's only remaining possible effect is to make an agent perform a non-maximizing action, whatever advantage originally accrued to adopting the disposition, it now maximizes to adopt one that will not require the agent to perform a non-maximizing action. That is, the constraining disposition is later made irrational precisely by the fact that it would only make the agent perform an action he prefers neither for its own sake nor for the sake of its consequences.

But can I answer Gauthier's challenge to explain why (as Christopher Morris put it) "the aim rationality gives us [must be] structurally identical to the decision rule it is rational to follow (in order to best achieve that aim)?"[25] The received view says rational agents aim at maximum expected utility and must follow the decision rule of maximizing expected utility at every decision node. Gauthier keeps the aim but changes the decision rule.

My answer follows from my internal critique of Gauthier. If a rational agent's aim is to maximize, then all that could rationalize a decision rule is that having it--i.e., tending to make decisions by it--maximizes. I claim to have shown that Gauthier has not come up with a situation in which making decisions by a maximization constraining rule is maximizing. Nor do I cheat or question-beg, for I do not merely repeat what is obvious--that any act of compliance with that rule will be non-maximizing--and then take a non-maximizing act to be irrational and so compliance with the rule to be irrational too. Rather, I grant (for the sake of argument) that if having the rule maximizes, choosing by it is rational. But in Gauthier's PD, having it is not maximizing by when one is to choose by it. I further conjecture that, since what makes it a non-maximizing decision rule to have then is that its sole effect then is to make its holder perform a non-maximizing action, if the aim of rational agents is maximization, such a rule can never be a rational guide to behaviour. For the only point to adopting it is that having it earlier puts one in a situation to profit (in terms of the aim of maximization) from breaking it (and so from having a disposition to break it) later. Thus, given the aim of maximization, only the decision rule that one should maximize in every choice has one's preferences giving one a reason to do what one does. And there should be no surprise here: we agree that the aim of the rational agent is to maximize; to maximize is simply to make as likely as possible the conditions for which one has preferences in the order in which one prefers them. In Gauthier's (intended) alternative, one's preferences only give one a reason to have the dispositions one has. And even there, as I show in my internal critique of Gauthier, the account collapses into mine. So in spite of Gauthier's larger theoretical claim, he unwittingly identifies rationality with maximization in every choice.

So, consider again Gauthier's claim that one would in fact do better by what one prefers, if one is (in effect) inclined to choose to do not what one prefers for its own sake or for its power to bring about something one prefers, nor what would make one better-off independently of what the other agent chooses, but what would make one best off, limited by what would also make him as best off. Cleaving to this principle is not

[25] From his referee's report.

now justified because one does not, unless one changes one's preferences, prefer either to be helpful to another or to cleave to the principle. Unrevised, one's preferences thus conflict with Gauthier's principle.

Even to make his proposal work, then, one must first revise one's preferences so that, all in, one prefers acting on the principle over minimizing one's jail time. But then one does not cease to be a maximizer. One only ceases to be selfish and morally unprincipled. So it is an acceptable variant on my solution to acquire a preference to conform to a principle like Gauthier's with just those with similar preferences. But then one would act on it not because one has it in a way that leaves one's preference-functions unchanged, but because one now prefers action on the principle to attainment of other outcomes.

In summary, a rational agent's principles are those of the maximizing conception of rationality. The continuingly rational agent does not revise his principles. Rather, the principles of choice making him rational require him to revise his preferences as a means of maximizing their satisfaction. Thereafter the maximization principle recommends something different, as his preferences are now different, which is why he then Co-operates.

Now Gauthier says that rational agents, as characterized by the received view (and, presumably, then, as characterized in my extension of it), cannot act from principle, only directly from their preferences. Yet on one hand, I have rationality as being the way one chooses, given one's preferences and beliefs. But then on the other hand, I have as morally principled persons those with a preference to conform to some moral principle. How can I have it both ways? How can I have a principle serve both as a way of choosing, given preferences, and as the object of a preference?[26]

Answer: Rational and moral principles are different kinds of things. The principles of rationality describe the relation between reasons and choices (or better, I think, express what it is for one's preferences to be reasons for choices, given beliefs, or reveal one's preferences in their capacity as reasons for choices). But moral principles are kinds of reasons. When one has internalized a moral principle--acquired a preference to choose conformably to it--one has acquired a reason for action. One has a reason to choose morally just if the rational expression of one's preferences given one's beliefs must consist in a moral action, e.g., in refraining from exploiting another agent. One has a reason to become moral (i.e., to acquire as reasons for action reasons which would rationalize one in performing moral actions), just if it is evident that one's current reasons qua aims are best served by the having of different (here, moral) reasons/aims.

There is no such thing as being moral simpliciter, i.e., independently of rationality. For being moral, surely, is not just a matter of behaving a certain way, e.g., non-exploitatively; it is a matter of so behaving for moral reasons. Moral reasons are just ones only moral behaviours could express. But behaving from moral reasons has in common with behaving from reasons in general, that it is maximizing, given one's reasons. Thus the moral are necessarily maximizers so far as they are genuine agents; what distinguishes them from non-moral agents is the aims on which they maximize.

[26] The question is, again, from Morris.

Arguments for the rationality of morality are therefore arguments that one ought to have certain aims, not that one ought to choose a certain way (e.g., non-maximizingly), given one's (possibly non-moral) aims. So choosing rationally is a way of choosing (to be contrasted with choosing non-rationally or irrationally); choosing morally is not a separate way of choosing. It is choosing rationally from moral aims.

11. Rationality and Alternative Conative Attitudes

It appears one must take the preference-revision solution if the original problem is what to do as a continuingly rational maximizer, given normal PD preferences. But could not the problem be given with a different ontology of conative attitudes? What if there were preferenceless beings governed entirely by brute dispositions to choose, or by rules (e.g., perfect Kantians, or program-following machines)?

The form of our solution still applies. If the question of the rationality of such beings and of their choices is even to arise, such beings would have to have the rules or dispositions structured in some way that admitted of an ordering--e.g., they would have to be disposed to maximize value _x_ first, _y_ second, etc. For we need a normative structure to serve as a basis for the normative evaluation of the instrumental rationality of their choices; we need something to distinguish the causes of their behaviours _simpliciter_ from the causes of their _appropriate_ behaviours. Thus we must imagine such agents having some standard by which to measure the appropriateness of their behaviours, something analogous to an _aim_ for behaviour. In the case of preferences, the aims of behaviours are the conditions which are the objects of preferences. For rules, the aims would be the states of affairs (outcomes) resulting from compliance with the rules; for brute dispositions, the consequences of behaviour successfully expressive of the disposition. Rational actions for such agents would then be actions which reflect or express these orderings. It may, then, sometimes be rational for them to revise their orderings--e.g., it may help them to maximize value _x_ that they become disposed to maximize value _y_ first. We might then inquire as to the rationality of their actions, given, say, their initial set of dispositions; and inquire as to the rationality of their adopting and choosing from a new set of dispositions as a way of better implementing a given set.

Imagine beings whose conations consisted of brute dispositions to follow certain rules. And suppose that their dominant initial rule is that one should spend as little time in jail as possible. As classically rational agents, their actions would have to conform to a principle to the effect that one ought to maximize conformity to one's current rules. Dominance reasoning would normally have them Defect, for only their third-best outcomes. But now we can apply my argument, but in terms of their conative psychology: they could more perfectly conform to their original rule by revising their rule-ordering, acquiring the rule that one should choose so as to minimize one's individual jail time, except that one should Co-operate with anyone holding as a higher rule that one should partake in joint Co-operation unless the other has as his highest rule that one ought always to Defect or always to Co-operate. Their newly ordered rules would then have them Co-operating with similar agents, for their originally second-best outcome.

Apparently, then, rationality is a function from conative attitudes to choices given beliefs, of which attitudes preferences are but one type. We might conclude that rationality requires one always to act on well-ordered attitudes that are the same in type: one cannot rationally justify action on dispositions, given one's preferences; or on preferences, given one's rules; or on rules for choice, given dispositions to choose, etc.--only on preferences rationally adopted, given initial preferences; rules, given initial rules, etc.

But one *can* rationalize adopting a heterogeneous conative set--e.g., adopting dispositions given preferences. It is just that one is only rational in *acting* from, say, the disposition, if doing so still expresses (maximizes on) one's preferences; for if the preferences determine selection of disposition (by way of determining the aims of appropriate behaviour, and so by setting a standard for a good disposition as one that would cause behaviours appropriate to the achieving of those aims), and if they survive after its adoption, it is they which ultimately rationalize the action from the disposition, if it is, indeed, a rational action. The disposition is just an intermediary. Basketball players, for example, inculcate in themselves dispositions and reflexes which thereafter operate automatically, without immediate (and perhaps psycho-kinetically clumsy) belief-desire antecedents. But their actions are still rational because those reflexes continue to serve the players' aims: they help them win.

I have argued that if we are inquiring into the rationality of principles, dispositions, plans, etc., given one's ordered preferences, then a disposition, plan, etc., is only rational if it causes choices which maximize on one's preferences. And a choice caused by a disposition, plan, etc., is rational only if it is a maximizing choice, not if it expresses a constraint on maximization. But we have also seen that we can imagine rational beings with no preferences, but with, say, internalized rules or brute dispositions; and that we can speak of the rationality of their choices of actions and of further dispositions or rules as bases for rational actions, on analogy with the maximization structure for preferences.

But surely we can also imagine beings (e.g., ourselves) with a heterogenous psychology, one containing both preferences and, say, brute dispositions. Does it not make sense to ask what it is rational for such a being to do, given its preferences and dispositions? If so, is it not possible for it rationally to alter its dispositions as a way of maximizing on its preferences--with the result that different things would now be rational, given those dispositions, e.g., maximization-constrained actions?

The answer to both questions is "No". Dispositions, plans, reflexes, internalized rules, principles of choice, etc., can be thought to exist in one's psychology in either of two ways. They either exist as mere causes of behaviour separate from reasons for behaviour, or they exist as themselves reasons for behaviour. If the former, we can speak of the rationality of a disposition, rule, etc., but we must then mean that behaving according to the rule results in behaviour which is maximizing, given one's preferences. If the latter, if these things are imagined to exist in something's psychology as *reasons* (as is the case if we imagine a machine with a set of programmed instructions, or a human being disciplined to choose as, say, a perfect Kantian), then we must imagine these things existing in its psychology in some way that is ordered. If these things exhausted its psychology, its rational actions would then be ones which reflected or expressed this ordering, i.e., ones which maximized, given the ordering.

But it makes no sense to ask what it is rational to do, given a set of ordered preferences _and_ a set of independently and differently ordered dispositions (or internalized rules or principles of choice, etc.) conceived independently of preferences, but yet still as reasons competitive with or supplemental to preferences. For to treat dispositions as reasons for action, the dispositions must be taken as ordered in a way analogous to the way preferences so order outcomes as to make a function from outcomes to degrees of satisfaction. This is functionally equivalent to a preference ranking, so that to imagine a heterogenous psychology, one with an ordered set of preferences and an independently and differently ordered set of dispositions, is to imagine a conflicted psychology. But rational actions can only be ones which reflect an internally consistent ordering. Thus since questions of rationality can only arise given a consistent ordering of conative attitudes, and since we have seen that rational choice from such an ordering is always maximizing choice, it follows that rational actions are always some analogue of maximizing actions. The maximization principle is now seen to give the structural form of all possible conclusive reasons for action, a structure violated by Gauthier's conception of constrained maximization. Thus, while it is possible rationally to choose dispositions and plans, and rationally to act from them, the resulting actions must always be maximizing actions, maximizing by some single rationalizing conative standard, e.g., that of a consistently ordered set of preferences.

Returning to the PD and generalizing, then, one initially conates (in some way) to minimize jail time. Dominance reasoning normally rationalizes acquisition of a conation for Defection. But that dooms one to one's third-best outcome with similar agents. Seeing this, one acquires a conation ordering for minimizing jail time except where one can Co-operate with those appropriately conating mutually Co-operative outcomes. Thereafter, one chooses action on the new conation set over action on the original one for minimal jail time _simpliciter_, since the latter is no longer part of one's psychology.

It does not matter which revision one makes in the psychological type of one's conation-set, provided one acquires conations, expression of which cannot be maximized except by Co-operating with those whose conation set's expression requires same. To get Co-operation from as many species of agent as possible, one should acquire the most conation-type-tolerant, properly ordered conation-set possible; one should prefer the Co-operative outcome with any agent whose conations would make him choose _as if_ he had the preferences argued for in Section 6, above.[27]

What if one faces agents who cannot modify their preferences, only their dispositions, and who have so modified them that they Co-operate with just the disposition-modified, remaining Defectors with those merely preference-modified? One must revise one's conation-type, not just its ordering. A preference maximizer is here justified in becoming a preferenceless ordered disposition-maximizer. (Whether he _can do_ this is another matter.) The prisoners could face a co-ordination problem in agreeing

[27] My thanks to Terry Tomkow for discussion on this point.

on a kind of behaviour-selecting gadget to adopt to move each to Co-operate when it is recognized in the other. But this is not their original problem, where each had an interest in being able to do something the other does not do, and in having attitudes towards actions allowing them to make choices he may not make. Here, it is in their interest to do the same things, and to have the same basis for choice, or ones sufficiently co-ordinating as to facilitate Co-operation; that apart, ceteris paribus, it is rationally indifferent which basis.[28]

12. Are Preferences Really Different From Other Conations?

These observations, however, raise another worry.[29] All conation-types--preferences, conceptions of rational principles of choice, causal mechanisms, internalized rules, brute dispositions to choose, etc.--function to select actions, just as preferences do in my solution. Yet, presumably, all of them work as behaviour-inducing mechanisms, causing a behaviour, just as merely mechanical dispositions do. So how can behaviours from preferences be rational, free, and voluntary actions, but not behaviours from rules, reflexes, etc.? And how then can I favour my solution to Gauthier's? If one acts voluntarily, etc., in complying with a preference, why not in acting from a disposition, and if not from a disposition, why from a preference? And if rationality requires me to abandon my CMD after it makes another Co-operate, why not also my Co-operative preferences? So are not preferences merely non-rational, blind causes of behaviour, just like mere dispositions?

No. To see this, consider a conation-set originally constituted by a rational psychology of preferences. It is logically possible for behaviour not to reveal any such preference, and for action not to reveal their balance (not to maximize), but not for rational behaviour to fail to reveal the balance of preferences. People sometimes act irrationally--by acting from irrational beliefs or preferences, or by acting irrationally from rational ones. Excepting one's initial preferences (in acquiring which, of course, instrumental rationality cannot figure), a new preference for behaviour is rational only if concludable from a practical syllogism, taking descriptions of antecedently held preferences and beliefs as premises. Likewise for actions.

I assume preferences and beliefs are identifiable independently of any given behaviour. A behaviour is rational only where caused by a preference, on balance, for its occurrence, or if it conduces to attaining some condition one prefers, on balance.

Both brute dispositions (as an example of a conative attitude-type alternative to preferences) and preferences may be mechanisms, and both may function to select behaviours (the difference is that while both are functions from beliefs to choices, preferences are also functions from outcomes to degrees of satisfaction, if it is preferences

[28] Compare also, Kurt Baier, "Rationality and Morality," Erkenntnis, 11 (1977), p. 213. My thanks to Terry Tomkow for help with this section.

[29] Thanks again to Peter Danielson for the following problems.

which are the reason-giving conations in the rational psychology in question), but they only do so rationally when they select ones *on balance* preferred, ones rationalized by the balance of preferences. This fits our discovery that rational action only requires a certain structural relation between one's conative attitudes and one's behaviours: it does not care which conative kinds populate one's psychology, only that one's behaviour accord with their ordered balance. Perhaps the distinction among types of attitudes is in name only; *qua* functions determining behaviours, they are indistinguishable. And perhaps we can only represent them as having a balance by ranking them as if they could bear a utility value. But Gauthier's solution is worrisome because he expressly conceives CMDs *not* to reflect the *balance of concurrent* preferences (only that of *earlier* ones); thus they are not rationalized by them. But only a balance of preferences (or of conative attitudes in general) can *rationalize* behaviours, make them rational actions. Conations other than preferences (whether other in name only or truly other), however, can, by premising behaviours, make for rational actions, if they form part of a conative psychology with a consistent ordering.

Another difference between CMDs *qua* irrevocable non-rationalizing mechanisms, and *qua* on balance preferences to Co-operate with those with appropriate conative functions: with the former, when one gets to go second in a sequential PD after having one's character assessed, one does not revoke the disposition only because one cannot, even though one has every reason to revoke it in one's original and surviving preference-set. With the latter, one has no reason to revoke, for one's conations have so changed that on balance they are now best served by Co-operation. There is no surviving set of conations such that one is "stuck" with a proclivity to Co-operate which one somehow wishes one did not have. One is now a happy and willing Co-operator.

Even if there can be a rational conative economy of things other than preferences, rational actions must still reflect their balance. Thus, in according dispositions a conative status comparable to preferences, one faces a choice: either the CMD is a balance of current such "preferences" favouring Co-operation, or it is a "preference" for action discordant with that balance. If the former, Gauthier's account is mine, and he should not say that Co-operation from a CMD is constrained, for it is just maximizing behaviour. But if dispositions are preferences, and if he conceives the CMD as somehow electing a behaviour without reflecting the balance of preferences, the resulting Co-operation is not rational action.

Inconsistently ordered conative attitudes must be avoided in a heterogeneous but rational conative psychology. For suppose I *prefer* to do *x*, but *believe in the rule* that one ought *not* to do *x*. What is it rational to do? Unless these attitudes admit to an ordering, that is just not clear. It is, perhaps, hard to say by what index preferences and beliefs about rules or principles of choice could relate as ordered. But surely to describe behaviour from these attitudes as constrained, as Gauthier does, is just to deny that they are well-ordered. And as we have seen, it is the *sine qua non* of rational choice to express the balance of ordered conative attitudes. Indeed, how else to represent one's conative state as affording a conclusive reason for action? To so represent is to say that all the vectors affecting behaviour sum to the behaviour chosen, sum not just causally, but justificationally.

Thus the rational agent is one who, whatever may be his conative attitudes, whether preferences, conceptions of rational principles for choice, etc., always chooses so as to express their balance. For to act rationally is just to act as one is on balance inclined to act, or in ways one believes most likely to achieve outcomes which one, on balance, is inclined to achieve. So suppose an agent faces a situation where, in order to maximize on the balance of his conative attitudes, he must become inclined to make choices which would not express that balance: to continue to be rational, he must first revise that balance so that it constitutes the required inclination, and then make the choices which would express the revised balance. We then have as just a special case the agent whose initial and later motivations are preferences: he must rationally revise their balance in agreeing to be moral for his later moral compliance to be rational.[30]

[30] I have here tried to elaborate on the proposal I made in "Two Gauthiers?" and "Libertarian Agency and Rational Morality," and to meet some of the objections I have received to it in correspondence and conversation. I have not had room to address objections to the very idea of a practically motivated revision in preferences. Does it make sense? Can agents revise their preferences? Is doing so consistent with all standards of rationality? with a conception of value as objective? I think the answer to these questions (except maybe the last) is "Yes"; but here I can only refer the reader to my "Preference Revision," where I defend this position in detail.