Two Gauthiers?*

DUNCAN MACINTOSH *Dalhousie University*

Introduction

In outline, at least, the major argument of David Gauthier's, *Morals By Agreement*[1] is quite clear: it is rational to acquire a disposition to moral behaviour with those similarly disposed because having it gives one the highest practically attainable utility in interactions with them. Once one has acquired such a disposition, a disposition which constrains one's individual maximization of the satisfaction of one's preferences (i.e., a Constrained Maximizer or "CM" disposition), it is rational to act morally towards similar agents. He thus takes himself to have constructed from instrumental rationality alone, a justification of voluntary compliance with moral principles. He appears to have shown that it is instrumentally rational for free and uncoerced agents to perform moral actions.

Still, many of the things Gauthier says are puzzling. In this paper I show that on Gauthier's own conception of practical rationality, moral choice and action cannot be represented both as voluntary, and as

* This paper began as a response to a talk given by Richmond Campbell to the philosophy colloquium at Dalhousie University in 1986, now in print as Campbell, 1988a. Also relevant is his 1988b. I am grateful to him for help with the literature, for searching criticism in discussion, and for detailed comments on an earlier draft. I have benefitted as well from dialogue with Neera Badhwar, Robert Bright, Robert Martin, Victoria McGeer, Geoffrey Sayre-McCord, Terrance Tomkow, Kadri Vihvelhin, Sheldon Wein, and from a quick conversation with David Gauthier. The referees for *Dialogue*, Wayne Sumner and Peter Danielson, offered very useful advice, and Danielson asked several important and difficult questions which, I think, require more extensive treatment than I have had room to give them here. I am especially indebted to Julia Colterjohn, who proved an invaluable foil throughout, and who read and offered comments on earlier versions. My thanks to the Killam Trust of Dalhousie University, whose post-doctoral fellowship support I enjoyed during the spring and summer of 1986.

1 All references are to Gauthier (1986), unless otherwise indicated.

issuing from a constraint on individual expected utility maximization. In resolving this tension, I consider two possible interpretations of what dispositions to moral behaviour might consist in. On the first, they are rationally acquired, irrevocable psychological mechanisms which determine but do not rationalize moral conduct. On the second, they are rationally acquired balances of preferences for behaving morally, given which preferences moral behaviour is rationally justified as straightforward individual expected utility maximization. I suggest that the latter interpretation fits best with the bulk of Gauthier's claims, and with the spirit of his argument, and I attempt a reconstruction of this theory from Gauthier's own writings.

1. Gauthier's Solution to the Prisoner's Dilemma

My discontents crystallize on Gauthier's treatment of the one-shot Prisoner's Dilemma (hereafter, "PD"). PDs are games where agents will each get a certain utility depending only on how they both have chosen in the game, and on their preferences regarding the material consequences of their combined choices. They each have a choice of co-operating or defecting. Each person's choice is presumed to be logically and causally independent of the other's. Each is known by each to be rational, and to prefer the material consequences of combinations of their choices in the following, decreasing order: Unilateral defection, mutual co-operation, mutual defection, unilateral co-operation. Thus an agent gets most utility from the material consequences of defecting where the other has co-operated, second most where both co-operate, third most where neither co-operates, and least by co-operating where the other defects. If he defects and the other co-operates he gets his best outcome, while if he had defected, he would only have gotten his second best. If he defects and the other defects, he gets his third best outcome, while if he had co-operated, he would only have gotten his fourth best. Thus, no matter how the other chooses, he does best by defection. This form of reasoning in a PD, dominance reasoning, seems to show that if rationality consists in individual expected utility maximization (hereafter "E-maximization") it is always rational, where one's choices cannot affect the choices of another, to defect; that is the dominant choice. Perversely enough though, the best this can yield agents competing with other rational agents is their third best payoff. If only both agents would make and keep agreements to co-operate, both could get their second-best payoff. Gauthier thinks that in light of this, rational agents should dispose themselves to constrain their maximizing behaviour so that they will co-operate with those they recognize to be similarly constrained; when agents recognize this constraint in each other, they will co-operate from the disposition, thus achieving their second-best payoff. He thinks co-operation under such a constraint consists in rational action. So far as

morality is conceived as a system of rules the following of which makes everyone better off than following any other rules, Gauthier thinks he has shown that it is rational to agree to follow such rules, and to keep that agreement, even if one could do better individually by breaking the rules—by defecting where others co-operate, thus attaining one's first rather than second-best outcome.

2. The Problem: How Can One Rationally, Freely, and Voluntarily Perform a Non-Maximizing Action From a CM Disposition?

Rational action is defined in the tradition, and by Gauthier himself, as the maximization of the satisfaction of present preferences. As Gauthier says, "[rational] choice maximizes preference fulfillment given belief" (30); "rational choice must be directed to the maximal fulfillment of our present considered preferences" (37). But standard dominance reasoning demonstrates that this dictates defection. Yet Gauthier thinks it is rational to constrain one's maximizing behaviour and co-operate. He thinks this will not only be rational, but totally free and voluntary behaviour.

So my first puzzle, then, is this: how can behaviour from a constraining disposition count as rational and voluntary action, given that it is not E-maximizing? I think that whether Gauthier has a satisfactory answer to this question depends on what he conceives CM dispositions to be. Unfortunately, his writings are systematically ambiguous on this point. They support two possible interpretations.

Mechanism Interpretation: CM dispositions are rationally self-chosen permanent mechanisms of some sort. Perhaps they are hard-wired psychological traits, hypnotically induced compulsions, socialized behavioural tendencies, strong habits, deadman switches, whatever. After they are acquired, they subsequently force their bearer's compliance with moral principles in interactions with those with similar mechanisms, independently of the agent's preferences at the time the mechanisms kick in.

Preference Interpretation: CM dispositions are preference-sets rationally revised to incorporate an intrinsic preference for moral *conduct* or *choice* with those with a similar preference. This preference is strong enough that its satisfaction, together with the satisfaction got from the material consequences of attaining the mutually co-operative outcome in the PD, would outweigh the utility attaching to the consequences of successful unilateral defection, given also the old and surviving preference for that *outcome* over all the others. In other words, to the original ordered preferences for outcomes, is added a higher-ranked preference for co-operative types of choice with those with similar preferences, independent of the consequences of outcomes to the issuing of which the choice will contribute. After the acquisition of such

preferences, agents' preference-sets issue in co-operative choices because such choices, given such preferences, would be E-maximizing.

I think that the preference interpretation is in fact the philosophically correct picture of things. But Gauthier says much to indicate that he does not hold it. And I think that no matter which interpretation we place on Gauthier's intention, consistency would require him to give up some of his other claims.

3. CM as a Mechanism Which Causes Behaviour Independently of Current Preferences

The problem is set in the following passage: "[a] constrained maximizer is conditionally disposed to co-operate in ways that, followed by all, would yield nearly optimal and fair outcomes, and *does* co-operate in such ways when she may actually expect to benefit" (177). The problem is that Gauthier does not say exactly *why* she *does* co-operate (at least not in that context). Certainly she would benefit *more* by defecting. Richmond Campbell (1988a) argues that she would only co-operate if she *had* to given her disposition; otherwise she would defect under the rationale of dominance reasoning. It seems to me then that the disposition merely *determines* co-operation, for it simply cannot rationalize it. Campbell thinks that a CM disposition must be a kind of stable reflex which makes one co-operate with others with a similar reflex. It would seem that it does this whether it would be E-maximizing to defect or not, and so whether one would at the time of co-operation on balance prefer to co-operate or not. The correctness of this understanding of what Gauthier means by CM is suggested by Gauthier's seeing it operating as a *constraint* on maximizing behaviour which affords others a guarantee that one will reciprocate their co-operation from such a disposition, in spite of it being E-maximizing for one to defect. Gauthier defends "the traditional conception of morality as a rational constraint on the pursuit of individual interest" (2). And there are many passages in which he seems to construe CM not as a preference, but as a brake on preference expression. (Emphases in what follows are mine):

> Duty *over-rides* advantage (2); we shall recognize the need for *restraining* each person's pursuit of her own utility (2); the rational principles for making choices *constrain* the actor pursuing his own interest in an impartial way (3); we shall ... undermine the force of the demand that rational choice reveal preference by showing that its scope may be restricted by ... a meta-choice, a choice about how to make choices. (79); [A] rational utility maximizer ... chooses, on utility maximizing grounds, *not to make further choices on those grounds* (158); These principles require a person to *refrain* from the direct pursuit of her maximum utility (168); [A] constrained maximizer may find herself required to act in such a way that she would have been better off had she *not* entered into co-operation (169); constrained maximization is *not* straightforward maximization in its most effective disguise (169).

This is not to say that Gauthier conceives co-operators *necessarily* as preferenceless or affectless beings so far as they co-operate. Some might well co-operate because they want to:

> In our argument, we have not appealed to any affective disposition; we do not want to weaken the position we must defeat, straightforward maximization, by supposing that persons are emotionally indisposed to follow it. But we may expect that in the process of socialization, efforts will be made to develop and cultivate each person's feelings so that, should she behave as an SM [i.e., as a straightforward maximizer, as someone who always defects], she will experience guilt. We may expect our affective capacities to be shaped by social practices in support of co-operative interaction (188).

But insofar as Gauthier does think preferences, affects, and so on *are* involved with having CM, he seems here to says that such preferences, if they arise at all, can arise *after* rational agreement to moral principles and *after* rational compliance with them, not necessarily after agreement to such principles but before or as a condition of compliance. Thus, CM, on the evidence of these passages, is not itself a preference or an affect.

Campbell, then, thinks Gauthier has, or should have, the mechanism interpretation in mind. And there is ample textual evidence to suppose that in fact he does.

4. Difficulties With the Constraining Mechanism Interpretation of the CM Disposition: A Vicious Dilemma For Gauthier

If CM is a permanent mechanism inducing co-operation in spite of defection being E-maximizing, and therefore, presumably being the more preferable action at the time of actual choice of action, it would seem that there is a crucial sense in which, at that time, one is *not* acting voluntarily—that is, from an immediate preference to so act—but merely at the behest of the mechanism. (See my 1987 on the extent to which behaviour from such a disposition might be counted as voluntary action in some non-obvious sense.) Further, it seems one is not acting rationally, because dominance reasoning shows that rational action on one's non-tuistic present preferences (i.e., one's preferences definable independently of the welfare or illfare of others) would be defection. (Again, for a more thorough discussion of this point, see my 1987.)

Yet it is clear that Gauthier thinks one co-operates *voluntarily* from a CM disposition. "[Internal moral constraints operate to ensure] ... those conditions under which individuals may rationally expect the degree of compliance from their fellows needed to elicit their own *voluntary* compliance" (164-165, my emphasis).

Were we to opt for the preference interpretation of what CM is, we could easily explain how co-operation from it can be regarded as voluntary compliance. There, CM is not a reflex which induces action defying present preferences, but is rather a revised set of preferences, ones such

that someone CM disposed in fact has an overriding preference for co-operation with those with a similar preference at the time he actually chooses co-operation. (Or perhaps better, he prefers, on balance, to co-operate with those with similar preferences.) Co-operation from such a preference (or balance of preferences) will be voluntary, since co-operation would then *consist* in doing what one on balance presently prefers to do.

But on the preference interpretation, co-operative choice and behaviour can no longer consist in *constrained* behaviour, since, with this new balance of preferences, one would prefer to co-operate, would get utility from co-operating, and indeed, co-operation would be the action which would, given those preferences, consist in E-maximizing action.

Thus, we have an apparently vicious dilemma: On the mechanism interpretation, Gauthier seems wrong to think of co-operation as voluntary, free, and rational action. On the preference interpretation, he seems wrong to think that it is constrained action. Moreover, these interpretational options seem logically exhaustive, for the agent's behaviours either express his preferences at the time of choice (as on the preference interpretation of CM) or they do not (as on the mechanism interpretation or some variant). In the remainder of this paper, I will consider whether CM as a permanent mechanism can be understood, in any sense that *Gauthier explicitly* countenances, to issue in rational, free, and voluntary co-operation. I will argue that by Gauthier's own standards it cannot. (For a review of some principled action-theoretic reasons why it cannot, whatever Gauthier's assumptions might imply about the matter, see my 1987.) I will then urge the interpretation of CM as a revised preference set incorporating an overriding preference for co-operation with those who share the preference, and show how to construct this interpretation from parts of some of the (confusing, and, perhaps, confused) later passages in Gauthier's book.

5. Difficulties With Gauthier's Defense of Constrained Maximization by the Dispositional Conception of Rationality

One reason for thinking Gauthier would count co-operation from CM *qua* mechanism as rational, free, and voluntary action, is that he sometimes seems to invoke a different conception of rationality, and, perhaps, of voluntary action, than is usually adopted in the tradition. He effectively says, at one point, that he conceives an action to be rational not if it is E-maximizing given present preferences and beliefs over any other action, but if it issues from a disposition the having of which is E-maximizing given present preferences and beliefs over the having of any other disposition. One might presume that, commensurately, an action would be voluntary not just if performed because it E-maximizes over any other action, but if caused by a disposition rationally chosen

because it E-maximizes over any other disposition. Given these definitions of rationality and voluntariness, the fact that the behaviour one's CM disposition induces is not E-maximizing over any other action, but is only the issuance from the disposition which E-maximizes over any other disposition, is no problem. It carries with it a new kind of voluntariness and a new kind of rationality, relative to which co-operation is both rational and voluntary.

But there are problems with this defense. Gauthier himself wishes to rationalize adoption of the CM disposition as the rational choice in a second-order choice problem, the problem of choosing not a behaviour, but a disposition to behave (or alternatively, the problem of choosing how to make all further choices, i.e., the problem of choosing a strategy, a policy, or a general intention). What disposition (or strategy) should one choose? He says the CM disposition. But it appears that he begins with the standard account of rationality in which it attaches to choices independently of dispositions, not to choices as issuing from dispositions. One proceeds from the former conception in choosing a disposition. This in itself may not be problematic. But in what consists rational action after one has chosen a disposition? Well, one might think that thereafter, it would be the action recommended by the disposition. But there is a difficulty here. Suppose we agree that the rationality of choices justifies adopting the CM disposition, and that the rationality of dispositions justifies co-operation thereafter. The difficulty is that the rationality of choices also justifies abandonment of the CM disposition after it has done its job of inducing others to co-operate (as when one gets to go second in a sequential PD), after which one could defect without penalty; at that point, if one can, one should abandon the CM disposition, adopt the disposition to always defect (the straightforward maximizing or SM disposition), and then, when the rationality of dispositions kicks in again, it mandates defection from the new SM disposition. (Perhaps the SM disposition just *is* the conception of the rationality of choices.) So if the principle of the rationality of dispositions is that it is rational to perform any action concordant with any disposition it is rational (because E-maximizing) to acquire, since it is rational (E-maximizing) to acquire the SM disposition after the CM disposition has done its work, the principle argues the rationality of defection, not of co-operation. (Mark Vorobej, in his 1986, makes similar objections to Gauthier's view that nuclear retaliation is rational given an originally maximizing intention or disposition to retaliate. I defend Gauthier on retaliation, much as I am about to do here, in my 1988b.) So why does Gauthier think initial choice and possession of CM rationalizes subsequent co-operation?

I think the only explanation is that he thinks acquiring the CM disposition is one's last-ever choice using the principle of the rationality of choices. The disposition is conceived as permanent and irrevocable. But

what rationalizes its permanence? What makes it continuingly rational in the face of a change in situation such that if one could abandon it and adopt a new, SM disposition, one would be better off?

Well, one might think that its permanence is rationalized by the fact that everyone knows that unless the disposition is permanent—unless it would prevent choices from preferences formed on the lines of dominance reasoning at the moment of actual independent choosing—possession of the disposition affords no guarantee of reciprocal co-operation. There would then be no reason to adopt the disposition, and even once adopted, it would make no difference to anyone's actual behaviour; they would all abandon the disposition and defect at the moment of choice, both to protect themselves, and to individually E-maximize. So it is apparently rational to adopt the disposition only and precisely because of its presumed permanence, and therefore, because of its assured efficacy as a determinant of subsequent behaviour.

But this does not prove co-operative behaviour issuing from such a disposition would be free, rational, voluntary action. It only proves that whether those behaviours have those properties or not, it may be rational, in selecting a disposition, to want everyone to have and continue to have, a permanent one, one guaranteed to have those issuances. This is to stave off otherwise inevitable mutual defection. But this does not change the fact that if one could escape the disposition, rationally, one ought to. The shared knowledge that if the dispositions were known to be revocable they would not afford an assurance of mutual co-operation, gives agents excellent instrumental reason to prefer, at the time of disposition adoption, that their dispositions bind and be known to bind their behaviours. But it does not follow that behaviours issuing from those dispositions are rational. Only that whether rational or not, they issue from dispositions welcome at the time of their adoption. At that time one is glad to have the option of being forced subsequently to behave irrationally with others similarly configured, if the alternative is mutual defection. How one feels about the disposition after it has done the work of inducing others to co-operate, however, is quite another matter. And dominance reasoning suggests that at that later time, one should experience a chafing at the bit.

Act and even disposition rationality as originally conceived, then, mandate the abandonment of the CM disposition by the time one actually comes to choose among actions, so *they* cannot be what rationalizes keeping it *by that later time*. Yet if *nothing* rationalizes keeping the CM disposition, then co-operation from it only because it is causally efficacious as a constraint on preferences and psychologically irrevocable, will be explicable behaviour, but not fully rational, voluntary or free action. It is difficult to see in what sense one is thereafter acting ration-

ally and voluntarily if one co-operates. Better to say that one has rationally causally configured oneself to behave irrationally.

Oddly enough, by Gauthier's own account, an action is rational only if it maximizes satisfaction of present preferences given beliefs: "Utility is ... the measure of present preference ... of the self at a particular time; practical rationality is the maximization of utility and so the maximization of the satisfaction of present preferences" (343). By his own account, the dispositions of a rational actor issue in rational, co-operative choices (187). There is a tension here. For by his own account, in choosing to co-operate, one (somehow) adheres to the constraining disposition "in the face of one's knowledge that one is not choosing the maximizing action" (186). Then he says,

> we do not purport to give a utility-maximizing justification for specific choices of adherence to a joint strategy. Rather we explain those choices by a general disposition to choose fair, optimizing actions whenever possible, and this tendency is then given a utility-maximizing justification (189).

It appears, then, that the disposition makes one act against one's present preferences. It cannot, then, by his own standard of rational action, *be* rational action. And it is suggestive that he says the disposition explains, rather than rationalizes, co-operative choices. For surely this means that behaviour from the disposition is not rational action, but non-rational merely caused behaviour, even though it issues from a disposition the acquisition of which was, perhaps, rational. That co-operation is not itself voluntary free-rational action, then, seems the inevitable conclusion given the mechanism interpretation of CM and given the definitions of practical rationality Gauthier accepts.

But while there is, in Gauthier, no argument justifying the claim that the action caused by the disposition inherits the rationality of the choice of the disposition, he does make such a claim: "Our argument identifies practical rationality with utility-maximization at the level of dispositions to choose, and carries through the implications of that identification in assessing the rationality of particular choices" (187). Now, however sensible it is to speak of the rationality of acquiring a disposition *qua* permanent mechanism, it seems extremely problematic how we are to carry through this identification in assessing the rationality of specific co-operative choices, especially in view of the larger conception of act-rationality which I have just reported Gauthier as endorsing, and which I believe to be the correct account of practical rationality. He seems to be shifting his definition of practical rationality around to suit the needs of his argument. It is E-maximization for selection of permanent dispositions, and disposition-following for subsequent action. But no principled rationale is given for the shift, and indeed, depending on the time in question, it seems that the E-maximizing criterion for assessing the rationality of a permanent disposition rationalizes different

dispositions—the CM disposition for when one is having one's character evaluated, the SM disposition for the time of actual choice after character-evaluation is complete. It does not give uniform advice. Indeed, it gives unfollowable advice, for presumably after acquiring CM, one *can not*, in accordance with the same principle that justified acquiring *it*, now revoke it and acquire SM; CM must be irrevocable if there is to be any rational reason to adopt it in the first place.

Attempting to defend Gauthier here, Campbell thinks that the initial rational preferability of the stable CM disposition over any other disposition assures the rationality of co-operative choices issuing from CM, since, he thinks, according to Gauthier, rationality as E-maximization is applied first at the level of dispositions to choose, rather than at the level of first-order choices. Now, I am not sure what "applying rationality" first to one level and then to another means. Presumably it is a recommendation about how rational agents should proceed in making choices. In any case, apparently if we do not do this—if instead we applied the E-maximizing standard first to choices, then to dispositions, one's disposition being read off one's choices—rationality as E-maximization would leave one as with the SM disposition, one which, paradoxically enough, affords one less utility than the CM disposition since it leads to mutual defection rather than mutual co-operation; hardly utility maximizing. To keep the conception of instrumental rationality consistent and "self-supporting" (Gauthier's phrase), Campbell thinks rationality must attach first to dispositions, then to choices. (See Campbell, 1988a, 202-203).

However, I think we have the reverse paradox if we go Campbell's route (and Gauthier's, if Campbell's reading of him is correct); we end up with specific choices of actions which are not E-maximizing, though with dispositions which are (at least initially; as we have seen, it appears that eventually there are conflicting recommendations about which disposition to have).

All this seems to me to show that something went wrong, and, I will now argue that the only consistent way out is to acknowledge the rationality of preference revision, and of co-operative choice upon acquisition of an on balance conditional preference for co-operation. Here, instrumental rationality will, throughout, be simple maximization of satisfaction of preferences. This takes us then to a detailed consideration of the preference interpretation of the nature of the CM disposition.

6. The CM Disposition as a Revised Balance of Preferences Favouring Conditional Co-operation

The preference interpretation suggests a conciliation between the demands of a uniform practical rationality and the need to assure mutual

co-operation. It allows us to retain the classical conception of act rationality, and the attendant conceptions of voluntariness and of which kinds of behaviours count as genuine actions (rather than as *merely* causally explicable behaviours), yet rationalizes keeping the disposition, and makes clearly rational such co-operative actions as may issue from it.

Suppose, as per the preference interpretation, we construe the CM disposition as just a set of preferences revised to incorporate a preference for justice (or for the welfare of others, or for co-operation with co-operators) for its own sake, practical act-rationality then justifying co-operative choices as maximizing of utility—of individual preference satisfaction—given the new preferences. Gauthier's argument should be understood to proceed by rationalizing an alteration in one's preferences (one alters them because in doing so, one can guarantee oneself one's second rather than third-best original payoff in the outcome in interactions with those who have similarly revised their preferences); once changed, specific co-operative choices would be rationalized as expressing or revealing the new preferences in accordance with instrumental rationality conceived as maximizing utility by maximizing satisfaction of concurrent preferences. Someone who has genuinely been moved by Gauthier's reasoning would prefer on balance to co-operate with someone with a similar balance of preferences, and would have no further reason to abandon the preference. Thus, we can account for the permanence of the disposition without supposing that it somehow continues to govern and constrain the agent's behaviour in contradiction to his preferences. It *is* his preferences, and since he no longer prefers a more minimal jail sentence over co-operating with a co-operator, he has no reason to acquire a preference or disposition to defect.

How does this fit with Gauthier's recommendations about the order in which to apply evaluations of rationality? That we should first evaluate the rationality of dispositions, and then take an action for rational just if it would issue from a rationally held permanent disposition, must be a procedural proposal; it is a piece of advice for rational agents that tells them how to best maximize their individual expected utility. It is also a criterion of rationality; we can use it to assess the rationality of perfectly free and informed agents. If it is a correct proposal, then it must itself be rationally defensible by an appeal to the principles defining instrumental rationality. The difficulty with construing CM as a permanent mechanism with efficacy independent of concurrent preferences is that while its adoption is rationalizable as E-maximizing, acting on it is not, throwing into doubt the extent to which behaviour caused by it is rational, voluntary and free, and even rendering problematic the intelligibility of counting it as an action at all. But if we suppose CM is a preference set, suddenly the procedural proposal makes sense. In effect it enjoins us to notice that given our preferences for outcomes and our perceived situation, assuming our psychological characters are transparent—are know-

able by ourselves and others—the preferences for choices we may have can affect the likelihood of our attaining an originally more preferred outcome. For our having a preference for co-operation with those with a similar preference will induce someone with a similar preference to co-operate with us as part of *his* maximal preference-satisfaction. Noting that, there arises the question of which preferences for choices to have in the service of one's preferences for outcomes. And it turns out that *if* one had a preference for the choice of co-operation with those with a similar preference, one would give others with the preference a reason to co-operate with one (because they would then want to), because having the preference gives oneself a reason to co-operate with them (one would then want to), and, so far as people act rationally from those preferences—i.e., so far as they do what they most prefer to do, viz., co-operate with those inclined to co-operate—the result will be mutual co-operation. One will get the superior utility attaching to one's originally second rather than third-best outcome. Further, once one has acquired such a preference, the preference itself can be used thereafter to rationalize the choice of co-operation, since given the preference, that choice is the E-maximizing one. We thus get choices of preference which are E-maximizing, *and* choices of actions from those preferences which are also E-maximizing, so that both choices are justified by a single uniform standard, that of E-maximization. Why not choose actions independently of a prior choice of preference for choices, and solely with a view to maximizing from one's original preferences? Because that would not be E-maximizing. One could do better by one's original preferences by first assessing and adjusting one's preferences. Thereafter, since rational action must express the balance of concurrent preferences, and one's new concurrent preferences favour co-operating, they now justify co-operation as E-maximizing action.

Note: I am claiming that it is rational (because maximizing on one's original preferences) to unilaterally alter the balance of one's preferences so that one favours co-operating just with those who favour co-operation (otherwise favouring whatever would individually minimize one's jail time), and that it is then rational (because maximizing on one's new current balance of preferences) to actually choose co-operation with those with a similar balance of preferences. This should not be confused with the idea, due to E. McClennen (1985), that rational agents in a PD should simply *resolve* to co-operate, and then keep to that resolve in choosing. Nor should it be confused with Amartya Sen's (1974a) proposal (on one interpretation of his article at least; but for controversy on this, see the exchange in Sen, 1974b; Watkins, 1974; Gauthier, 177; and Baier, 1977) that they should act towards each other *as if* their first choice was to co-operate; nor with a variant on that proposal that they should acquire a second-order preference for ways of

acting on their original ordered preferences for outcomes and actions, a preference for acting co-operatively in spite of their preference that their individual jail time be minimal in the outcome. Finally, it should not be confused with a proposal like the one David Lewis (replying to Gauthier, 1984) considers and rejects while assessing the rationality of retaliating from a rationally adopted intention to retaliate (in Lewis, 1984, 153-154), viz., that rational agents should implant in themselves a preference to co-operate and then act on it. The difficulty with all of *these* proposals, is that they either leave the balance of the agent's preferences such that defection is E-maximizing, or they give one incoherently ordered preferences. Given a preference for minimum jail time, why keep to McClennen's proposed resolutions? Why act, per the first version of Sen, as if one preferred to co-operate when one does not? Why act, per the second version of Sen, on a second-order preference directly at odds with one's first-order preferences? Why act, per the proposal Lewis rejects, on a preference to co-operate discordant with the balance of one's preferences? (For a more detailed discussion of these proposals, see my 1988a.) This is a generic problem with solutions to the PD that do not rationalize alterations in the preference functions of the individuals in the game. It is the same problem that plagues those versions of Gauthier in which he is advocating that one acquire and act on a constraining disposition *qua* mechanism, or that one should acquire and act on a new principle of choice given preferences, namely the CM principle: Why act on a disposition or principle which has one doing less well by the satisfaction of one's preferences then if one just maximized? The crucial difference between these proposals and mine: I say one must revise the overall balance of one's preferences so that one prefers, above all else, to have co-operated with those who prefer to co-operate with conditional co-operators, and otherwise to have an individually minimum jail time. This will have one co-operating with conditional co-operators, otherwise defecting as usual. Note too that the other proposals (except Gauthier's) fail to rationalize unilateral self-alteration. What if the other agent does not resolve, or does not acquire a preference to act as if his first-order preferences favoured co-operation, etc.? Their proposed forms of self-alteration do not leave the agent safe from exploitation. Unilaterally acquiring a conditionally co-operative preference does safe-guard one from exploitation, however, since it only has one co-operating with agents who would find it rational to reciprocate. Finally, note that the preference for co-operation with conditional co-operators does not violate the requirement that the agents' choices be independent of each other. One rationally acquires the preference whether the other does or not; one co-operates from the preference just if the other has the preference (not just if the other co-operates—*that* would be a vicious dependence, each waiting interminably for the other's decision.)

7. Explaining Gauthier's Substantive Conclusions With the Preference Interpretation of the CM Disposition

Now I have just argued that it would make sense to hold that rationality licenses acquisition of a new affect or preference (on balance) for co-operation with those similarly inclined, and that co-operation is not fully rational, voluntary, and free action until those preferences are acquired by both agents. (Compare with 327-329.) This proposal is certainly not incompatible with the spirit of Gauthier's proposal, nor indeed, with what, elsewhere, he seems to explicitly say. Consider, again, some relevant passages from Gauthier: "it is rational to be disposed to constrain maximizing behaviour by *internalizing* moral principles to govern one's choices" (15). Much depends, of course, on what he means by "internalizing" here. That we should take internalizing moral principles to amount to revising one's preferences so that one prefers choosing morally is suggested by some later passages in Gauthier:

> The just person is disposed to comply with the requirements of the principle of minimax relative concession in interacting with those of his fellows whom he believes to be similarly disposed. [He] is fit for society because he has internalized the idea of mutual benefit, so that in choosing his course of action he gives primary consideration to the prospect of realizing the co-operative outcome. If he ... may reasonably expect to bring about an outcome that is both (nearly) fair and (nearly) optimal, then he chooses to do so; only if he may not reasonably expect this does he choose to maximize his own utility (157).

Now surely it would be mysterious in what sense he had internalized the idea of mutual benefit, if this is to be contrasted with maximizing his own utility; surely if one has internalized the former, trying to realize the co-operative outcome would maximize the latter. Elsewhere, contrasting his position with that of Hobbes, Gauthier writes, "Hobbes does not suppose that each man internalizes the right reason of the sovereign" (163). "[Hobbes'] ... egoistic psychology allows the internalization of no standard other than that of direct concern with individual preservation and contentment" (163). On Gauthier's account just the reverse is true. Agents *are* able to internalize a standard other than that of direct concern with individual preservation and contentment. This by itself does not prove that the internalization of the standard consists in the adoption of some new preferences for co-operative choices in interactions with those with a similar preference. But that interpretation *is* forced upon us by the following: Hobbes' agents are straightforward E-maximizers. Thus *their* internalized dispositions must just be preferences, since all E-maximizers do is maximize satisfaction of their individual preferences. This implies that in general, internalized standards or dispositions are the same kinds of things as preferences. Gauthier then, to be consistent, must acknowledge that that makes the CM disposition (or standard), once internalized, also just a set of preferences. We should thus construe co-operation with those similarly dis-

posed from the CM disposition as yet more E-maximizing; it merely happens that Gauthier's agents rationally acquire a preference for justice. Surely that makes them no less E-maximizers. For as he himself points out (7), being a rational maximizer is not necessarily maximizing expression of concern *for* self, but of concern *of* self (which may or may not be *for* self), and that may, surely, be concern of self for others' welfare, and/or for fairness in the distribution of utilities. Did he mean all along that individual utility maximization was just maximization of satisfaction of concerns *for* self? He definitely wishes to initially presuppose nothing more than such non-tuistic preferences in agents in trying to show that all instrumentally rational agents ought to be morally compliant. But since people, once they *have* acquired the appropriate preferences, surely *can* get utility from just or generous action, can be made happy by the happiness of others, etc., surely he did not mean to exclude this possibility in principle.

He says many other things which can, I think, only be explained by invoking the preference interpretation.

> A *constrained* [Gauthier's emphasis] maximizer ... seeks ... to maximize her utility, given not the strategies but the utilities of those with whom she interacts (167); [She] is ready to co-operate in ways that, if followed by all, would yield outcomes that she would find beneficial and not unfair, and she does co-operate should she expect an actual practice or activity to be beneficial (167); [honesty is to be treated] not as a [mere] policy [i.e., strategy], but as a disposition. Only the person truly disposed to honesty and justice may expect fully to realize their benefits, for only such a person may rationally be admitted to those mutually beneficial agreements ... that rest on honesty and justice; on voluntary compliance. But such a person is not able, given her disposition, to take advantage of the "exceptions": she rightly judges such conduct irrational (182).

We can now explain why someone would not take advantage of the opportunity for unpunished dishonesty in the breaking of agreements; she would not, because she would find doing so irrational, and she would find it irrational because cheating would go against her rationally acquired overriding preference (or better, her balance of preferences) to choose fairly in interactions with those with a similar preference. To be sure, one wonders in what sense the exceptions "afford opportunity for advantage" if one has really internalized dispositions *qua* preferences against cheating. Perhaps what Gauthier meant (or should have meant) was that, though cheating would enable one to better satisfy one's preference for an individually best outcome, it would involve going against one's preference for making co-operative choices, which preference is sufficiently strong as to override the preference for a better outcome. Thus construed, the disposition renders compliance with morality voluntary in the sense of being concordant with concurrent preferences given beliefs, i.e., in the sense of being E-maximizing given one's present, coherent, ordered, all-things-considered preferences.

8. Dealing With the Apparent Ambiguities In Gauthier's Theory

Perhaps the tension in Gauthier between the mechanism and preference accounts can be resolved this way: People beginning with no fellow-feeling or commitment to fairness can be brought to agree to moral principles, and can dispose themselves to conform to them. For those with a capacity for an affective morality (i.e., for those who are psychologically capable of acquiring a preference for fair conduct, or for others' welfare, if only they are given good reason to do so), Gauthier's arguments can be construed as rationally justifying acquisition of an affective regard or preference for others' welfare as well as for one's own, and/or for fair behaviour in interactions with those with similar preferences. For those who begin with an already present affective morality (i.e., for those who already prefer to co-operate with the similarly preferring), his argument can be seen as a rational reconstruction of a retrospective justification for that morality. If one at present lacks affective morality, and if one does not even have the psychological potential to acquire same, Gauthier's argument rationalizes acquisition of a permanent mechanism which will guarantee co-operation, not as rational, voluntary, free compliance, but as irrational or non-rational behaviour from a disposition it is rational to adopt. But his only scruple against initially taking CM for a preference, is that he is greatly concerned not to presuppose that agents begin with such an affective regard for others, or with such a preference for fair co-operation with those with similar preferences, in demonstrating the rationality of morality; affective morality is rather something the aptness of which he takes himself to have proved without assuming it, at least for ordinary people in ordinary circumstances, who have at least the capacity for affective morality, for beneficiently other-regarding choices, and/or for fairness-aiming choices. (See 327-328.)

His mistake is to think, if he does, that people who have not been or cannot be moved to an affective morality by Gauthier's rationale, can still rationally, voluntarily, freely co-operate. As Campbell has shown of persons free in the Libertarian sense, they can not. They could achieve co-operation only through causally stable co-operative mechanisms which would induce them to behave irrationally relative to their purely and permanently egoistic preferences. (Gauthier himself seems to recognize this when he says that Economic man, given by the ring of Gyges the power to deceive and escape all punishment, will lack any rationalization for fair behaviour. See Gauthier, chap. 10.) But if it is rational to alter one's preferences, preference-resistantly permanent mechanisms are not needed.

But are the preference and mechanism interpretations really different? Peter Danielson questioned this (in correspondence). On either view CM dispositions play the functional role of preferences in that they

select actions; but both do their work regardless of their rationales, like mechanisms. How then can I say that if CMs are preferences, co-operation is rational, free, and voluntary, but if they are mechanisms, not? What makes action from those "blind" mechanisms we call preferences so special? On either interpretation, do not CM dispositions have the same properties, and so the same virtues and vices?

My reply: even if we use functional considerations to conflate mechanisms with preferences, the CM disposition is still either an isolated preference/mechanism discordant with the balance of one's other preferences/mechanisms, or it is integrated with them and makes for a balance of preferences favouring co-operation with conditional co-operators. If the former, co-operation is still irrational, involuntary, and unfree by the standard of maximization on the balance of one's preferences; if the latter, co-operation is then rational, etc., but then Gauthier should not conceive it as constrained behaviour. It is, in fact, just more maximizing behaviour. In any case, I want to keep the distinction between behavioral causes which are reasons, and those which are not. One's behaviour is rational, voluntary, and free just if caused by and/or appropriate given the balance of one's present reasons, when judged by the E-maximization standard. In that case, one acts as one does because that is what one comes to want to do, all things considered. Not so when one behaves from a mere mechanism understood as discordant with the balance of one's current preferences. One is a slave if behaving from a mechanism not rationalized by the balance of concurrent preferences, but the master of one's destiny if one acts as one on balance prefers. This is why one is no slave if one co-operates from a revised balance of preferences; one is only doing what one then wants to do. But one is a slave if one co-operates from a *mere* mechanism; it makes you do something you do not want to be doing. (See my 1988a for more on this issue.)

Perhaps Gauthier's best formulation of his account on the preference-revision interpretation of his views, is this:

> Persons rationally recognize the constraints of morality as conditions of mutually beneficial co-operation. They then come to value participation in co-operative and shared activities that meet these constraints, and to take an interest in their fellow participants. And finally they come to value the morality that first appeared to them only as a rational constraint (338). A rational morality is contractarian. But this does not imply that it is of purely instrumental value to us. In relating morality to the provision of benefits that themselves involve no affective concern with others, we do not thereby impoverish the moral feelings of persons who have such concern. It is because we can give morality a rational basis that we can secure its affective hold (339).

In further support of the preference interpretation, we can note that Gauthier gives ample evidence of believing that people can rationally alter their preferences. In objecting to political and authoritarian solutions to the PD, Gauthier writes, "artificial justice, in adapting persons to institutions, fails adequately to accommodate the rational capacity of

persons to reflect self-critically on their preferences" (342). In explaining the agent's relation to his future selves, he claims that, "[t]he utility function that partially defines a particular individual is revisable, in part through the individual's own reflective activity" (342). Elsewhere, he says, relevantly: "Morals by agreement capture the understanding of economic man; they capture the affections of the liberal individual" (345); "An affective capacity for morality is needed if the constraints required by essential justice are to be *willingly* honoured" (348, my emphasis). Otherwise, he thinks, it would have to be sustained by socialization, and by enforced roles individuals would not be free to accept or reject.

> This process requires ... preferences and capacities ... to serve as inputs, and there is no threat to autonomy in the recognition that these inputs are not, at least initially, autonomously determined. What makes a being autonomous is his capacity to alter given preferences by a rational, self-critical, reflective procedure, not a capacity to produce preferences with no prior basis (349); We suppose that persons are soft-wired so that they may change their desires and aims (350); an essentially just society must be strengthened through the development of the affections and interests of the young in such a way that their mature concerns afford motivational reinforcement to the rational requirements of co-operation. Co-operative activity should be experienced as itself fulfilling. Socialization, then, should encourage persons to want to co-operate in those situations in which co-operation is otherwise mutually advantageous to them. The desire to co-operate in such circumstances will receive the reflective endorsement of reason; the justification of the essentially just society extends to the justification of the sociability that sustains and strengthens it, and so to the justification of the socialization that instills and encourages this sociability (351); An animal with the right to make promises must be able to commit itself, giving itself a reason for choice and action that overrides its usual concern with fulfilling its preferences. Such an animal is able to interact with its world in a new and distinctive way, which we have sought to capture in the conception of constrained maximization (355).

The great difficulty in interpreting Gauthier is his tendency to conflate truly voluntary compliance with compliance from a rationally adopted constraining disposition in his rhetoric, the passage from 350-355, sampled above, being a good case in point.

References

BAIER, KURT, 1977
"Rationality and Morality", *Erkenntnis* 11, 197-223.

CAMPBELL, RICHMOND, and LANNING SOWDEN, eds., 1985
Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem. Vancouver: University of British Columbia Press.

———, 1988a
"Moral Justification and Freedom", *The Journal of Philosophy* 85/4, April, 192-213.

———, 1988b
"Critical Study: Gauthier's Theory of Morals by Agreement", forthcoming in *The Philosophical Quarterly*.

GAUTHIER, DAVID, 1977
"Critical Notice" of Körner, ed., 1974, *Dialogue* 16/3, 510-518.

———, 1984
"Deterrence, Maximization, and Rationality", *Ethics* 94, 474-495.
———, 1986
Morals By Agreement. Oxford: Clarendon Press.
KÖRNER, STEPHAN, ed., 1974
Practical Reasoning. Oxford: Basil Blackwell.
LEWIS, DAVID, 1984
"Devil's Bargains and the Real World", in Maclean, ed., 141-154.
MACINTOSH, DUNCAN, 1987
"Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem". Manuscript.
———, 1988a
"Preference's Progress: A Post-Script to 'Two Gauthiers?' ". Manuscript.
———, 1988b
"Retaliation Rationalized". (Manuscript; presented to the *Canadian Philosophical Association*, May, 1988.)
MACLEAN, DOUGLAS, ed., 1984
The Security Gamble: Deterrence Dilemmas in the Nuclear Age. Totowa, NJ: Rowan and Allenheld.
MCCLENNAN, EDWARD, F., 1985
"Prisoner's Dilemma and Resolute Choice", in Campbell and Sowden, eds., 94-104.
SEN, AMARTYA, 1974a
"Choice, Orderings and Morality", in Körner, ed., 54-67.
———, 1974b
"Reply to Comments", in Körner, ed., 78-82.
———, 1977
"Rationality and Morality: A Reply", *Erkenntnis* 11, 225-232.
VOROBEJ, MARK, 1986
"Gauthier on Deterrence", *Dialogue* 25/3, 471-476.
WATKINS, J. W. N., 1974
"Comment: 'Self-Interest and Morality' ", in Körner, ed., 67-77.

CANADIAN JOURNAL OF PHILOSOPHY

SUPPLEMENTARY VOLUME 13, 1987

SCIENCE, MORALITY AND FEMINIST THEORY

Editors: **Marsha Hanen and Kai Nielsen**

Introduction: Toward Integration — **Marsha Hanen**

Two Aspects: Science and Morality
- Sex Inequality and Bias in Sex Differences Research — **Alison M. Jaggar**
- The Need for More Than Justice — **Annette C. Baier**

Critiques: Science, Ethics and Method
- The Philosophy of Ambivalence: Sandra Harding on *The Science Question in Feminism* — **Alison Wylie**
- Ascetic Intellectual Opportunities: Reply to Alison Wylie — **Sandra Harding**
- Beyond Caring: The De-Moralization of Gender — **Marilyn Friedman**
- Non-Contractual Society — **Virginia Held**
- Rawls and Ownership: The Forgotten Category of Reproductive Labour — **Sibyl Schwarzenbach**
- Ethics, Ideology, and Feminine Virtue — **John Exdell**
- Women and Moral Madness — **Kathryn Morgan**
- Moral Sanity or Who Killed Boy Staunton — **Steven Burns**
- Rescuing Womanly Virtues: Some Dangers of Moral Reclamation — **Barbara Houston**

Some Applications
- Feminist Ethics and In Vitro Fertilization — **Susan Sherwin**
- Surrogate Motherhood — **Christine Overall**

Selves and Integration
- Only Connect: The Place of Self-Knowledge in Ethics — **Sheila Mullett**
- A Feminist Aspect Therapy of the Self — **Ann Ferguson**
- Second Persons — **Lorraine Code**

Afterword: Feminist Theory - Some Twistings and Turnings — **Kai Nielsen**

This Supplementary Volume is free to individual and student subscribers to CJP Volume 17, 1987.

PRICE:
CDN $14.00 in Canada US $12.00 outside Canada. Please add $1.50 for postage and handling, and 50¢ for every further copy.

ORDER FROM:
The University of Calgary Press, LT1013, The University of Calgary, 2500 University Drive N.W., CALGARY, Alberta T2N 1N4 CANADA.

ISSN 0229-7051 ISBN 0-919491-13-8