RETALIATION RATIONALIZED: GAUTHIER'S SOLUTION TO THE DETERRENCE DILEMMA[1]

Duncan MacIntosh,
Dalhousie University

Published as: Duncan MacIntosh, "Retaliation Rationalized: Gauthier's Solution to the Deterrence Dilemma", Pacific Philosophical Quarterly, Vol. 72 No. 1 (1991), pp. 9-32.

1. Introduction

David Gauthier has perhaps done more than anyone to challenge the central dogma in the theory of instrumental rationality, that it is always rational to maximize one's individual expected utility. His arguments exploit situations where it maximizes for agents to intend to perform a certain action, but not to actually perform it. He thinks that since it maximizes to intend the actions, intending them is rational. An intention's rationality entails that of the act intended, even if it is non-maximizing. So it is not always rational to maximize.[2]

It sounds simple. But as Gregory Kavka has taught us, these scenarios abound in paradox. It is extremely controversial how it would be rational to act in them, extremely contentious what they show about rationality, intention, and action. Kavka's classic scenario (the Deterrence Dilemma, or DD)[3]: Unless you, a nuclear superpower, really intend to all-out retaliate if all-out attacked by an enemy power, and unless this will guide your later behavior so your enemy will fear to attack, you will likely be attacked. Your only relevant preference is that harms to all be minimal. Intending would likely minimize harms for it would likely deter. So it maximizes. But if you are attacked, it does not maximize to retaliate; that would only cause pointless further harms.

Such cases raise, first, the intention problem: can one rationally acquire an intention it maximizes to acquire, but not to act upon; can one rationally intend to cause what one disprefers? Second, the action problem: if one could rationally acquire the intention, could one rationally act on it? These tend to raise, third, the link problem: does the rationality of an intention stand with that of the action intended? Fourth, the determination problem: if so, which decides their fate?

Kavka thinks the situation paradoxical, the choices partly linked. There is reason to intend for that maximizes, reason not to act for that does not. The irrationality of acting makes it impossible to directly and rationally come to intend, for one can't rationally intend an irrational act. Even if one could intend, the reason not to act still makes acting irrational; it is not maximizing. But no matter what one does, one's rationality is tainted; either one's actions or one's intentions are irrational: If one refrains from intending, one irrationally allows the harms of an attack. If one intends, one intends to do something irrational. If one intends and acts, one does something irrational from an intention it was rational to have (if not one directly rationally acquirable). If one intends and doesn't act, one refrains from an irrational act, but fails to conform to an intention it was rational to (somehow) acquire.[4] Kavka sees a conflict between our criteria for the rationality of

persons, intentions, and actions. We say an intention is rational if the act intended is; an action, if it maximizes one's individual expected utility given one's preferences; a person, if his intentions and actions are rational. This is fine where intentions are parts of the actions intended, or have no other effect but to proximally cause actions. But where intentions are actions separate from those intended, or have separate effects, assessments of agents' rationality by the effects of their intentions conflict with assessments by the effects of the actions intended.

Gauthier tries to cut through these problems by suggesting that while rational agents must maximize, they need not always make maximizing choices. Sometimes it maximizes over-all if they adopt intentions, strategies, or policies later constraining them from making maximizing choices; in so adopting, they better serve their aims than by maximizing in every choice. Their actions are then rational if they express maximizing intentions. Thus that these actions aren't maximizing is neither an objection to intending to perform them and actually performing them, nor to an agent who both intends and acts on the intention being rational. Gauthier, then, thinks the DD shows an action needn't maximize to be rational; it need only express a disposition it maximized to adopt.[5]

Does Gauthier's argument successfully resolve these paradoxes? I have recently argued that it does not in the analogous case of the Prisoners' Dilemma (where it maximizes to intend to co-operate with someone who would reciprocate if one so intended, but where it does not maximize to actually co-operate). Mark Vorobej has recently argued that Gauthier's argument does not work for the DD. Our arguments have a common structure. Here's Vorobej's: Even if actions are rational not if maximizing but if expressive of a maximizing disposition, it is still irrational both to intend and to act. For while the maximizing disposition to have pre-attack is retaliatory, the maximizing one to have post-is non-retaliatory. It would dictate non-retaliation. So retaliation is irrational. But since it is only rational to intend rational acts, and since retaliation is irrational, so is intending it.

I have not been able to shake the worry that this argument is too quick. In this paper, I consider a variety of ways Gauthier might be defended. I draw on some of his own speculations about the nature of rationality, intentions, choices and actions, and about the relation between one's preferences and one's reasons for action. I consider the possibility of extending his theory of rationality into a new theory of action, one which initially seems to afford Gauthier a reply to Vorobej and myself. I conclude, however, that Gauthier's argument simply cannot be made to work on its own terms. I claim that not only doesn't the DD falsify the maximizing conception of practical rationality, but that Gauthier's is really the same theory in different words. I do, however, think it is rational both to intend to retaliate and to actually retaliate in the DD. How can I believe this if I think Gauthier's argument does not work? I think that Gauthier and others in the field have been coming at the problem with a false assumption.

Gauthier and his critics have assumed that one can dispute the nature of rationality and the rationality of intentions and actions, given fixed preferences. But I argue that what is really at issue here is the rationality of preferences. I think DDs show that it is sometimes irrational to prefer as one does, namely, where it would maximize on one's current preference to have different preferences. Once we see that it is rational for one's

preferences to change in DDs, we will see that intending to retaliate and acting on that intention are rational. Yet the classical doctrines that it is rational to maximize, and that one can only intend and perform maximizing actions, are preserved. But decision theory, game theory, and our understanding of classical rationality need overhauling. For rationality really tells us not how to choose just given our preferences, but what we should prefer given what we now prefer, and how we should choose given what we should prefer.

2. The Gauthier Arguments

Gauthier thinks it rational to intend and to act. His thinking runs as follows.[6] First, a preamble establishing the prima facie rationality of intending:

(G1) Were it irrational to intend, a free and rational agent could not advantage himself by deterrence, for he could not credibly threaten. But a rational action maximizes. Intending maximizes, so it must be rational.

Now Gauthier thinks an intention is rational only if the act intended is too. So he gives two transcendental arguments for the derivative rationality of acting:

(G2) It is impossible to rationally intend the irrational. It is possible and rationally obligatory to intend to retaliate. So retaliation must be rational.

(G3) Intending is rational only if it would deter. It would only deter if it would probably cause retaliation. It would only do that, if its agent stayed rational, free and well-informed, if retaliation were rational. Since it is rational to intend in a way that would induce retaliation, retaliation must be rational.

These sound like bad arguments. They have the form: it would advantage you were x true, so x must be true. Non sequitur. Moreover their conclusion is prima facie false. It is not maximizing to act. It is thus irrational to act. It is not possible to rationally intend an irrational act. It is thus impossible to rationally intend. Or, if it is somehow possible, it remains irrational to act. We may wish it were rational to intend and even to act, were that necessary to its being rational to intend. But that doesn't show either is rational.

Is this a fair reductio? Perhaps not. Perhaps it works against arguments for a proposition's truth, but not against ones for an action's rationality. Perhaps the very scope of the rational, unlike the true, is constructed from the concept of the maximizing. If action x maximizes, x is rational. If action y's being causable by x is necessary to x's being rational, y too must be rational. x maximizes, so y must be rational.

Allright; perhaps if x maximizes, x is rational. But it does not follow from x being rational only if it can cause y that if x is rational, so is y--only that y must be causable by x. x can be maximizing even if y isn't; thus x can be rational even if y isn't. Thus, x's rationality doesn't entail y's, even where x induces y. Moreover, if x was rational only if y was if x causes y, then if y is irrational, x is too, even if maximizing. Or perhaps we have conflicting standards of an action's rationality, vis., its being maximizing, vs. its being caused by something maximizing: what of things caused by maximizing things, but not themselves maximizing? What, in short, of Kavka's paradox?

Gauthier thinks this objection misconstrues what rationality requires of an agent in a DD. The objection has a false premise, that it is always irrational to perform a non-

maximizing action. Thus:

(G4) Classically, rational actions must maximize. But some possible actions involve revising one's nature. E.g., consider one's dispositions to choose. A classical or Straightforward Maximizer (SMer) is disposed to maximize in each choice. He can't rationally retaliate, for that is not maximizing. If rationally intending retaliation requires retaliation to be rational, he can't intend; nor then can he avail himself of a credible deterrent intention. But suppose instead that, in choosing actions, he first chose policies, intentions, strategies, whatever (call them "dispositions"), the having of which maximized. And suppose he tended to act on these (for otherwise, he couldn't advantage himself with them; no one would fear that he would act on them, and it might be senseless even to say he had them). It would advantage him to choose a retaliatory disposition, for it would be a credible deterrent if he could be relied upon to act on it, as he could if he tended to act on his dispositions. Now suppose that he is, initially, an SMer, but has the option to be a Disposition Maximizer (DMer), one who always first chooses the maximizing disposition, then chooses from it. It would be maximizing (because deterring) and so rational for him to become a DMer with a retaliatory disposition. Thereafter, Gauthier thinks he will retaliate.

This may establish the rationality of intending or being disposed to retaliate, and of becoming someone who acts on his dispositions. But does it make so acting rational? Or does it merely prove it rational to adopt dispositions to cause what are really irrational behaviors? Gauthier thinks that to say it is rational to intend is to say it is rational to act. So it would be inconsistent to grant the rationality of intending to retaliate but not of retaliation, like saying it is both right and not right to retaliate. So if we grant it is rational to intend, we must grant that it is rational to act (provided nothing changes about one's circumstances, apart from the conditionally anticipated fact of one's intention failing to achieve the end for which it was adopted); indeed, he thinks an action inherits rationality from that of the disposition or intention to so act. Call this (following Mark Vorobej[7], who follows Derek Parfit) the Inheritance Principle (IP).

But surely someone who now acts on an old policy is irrational if implementing it isn't maximizing? Gauthier thinks not. At worst, he is acting in an "irrational manner," i.e., without further deliberation. He needn't reconsider what to do; he has a policy, and he executes it. This is rational were it rational to adopt a policy to act non-deliberatively, as it was.

Gauthier's critics will accuse him of bad faith on the rationality of actions[8]; he calls actions rational even though dispreferred (or with dispreferred consequences), if dictated by an expedient disposition. He may reply that his critics are rascals for counting inutile intentions as rational; intending not to retaliate courts attack, even if it saves the agent from intending a dispreferred action, or one with a dispreferred consequence. Gauthier thinks then that classical rationality must acknowledge the rationality of intending, this entailing that of acting. Thus non-maximizing actions can be rational.

3. Vorobej's Replies to Gauthier

Mark Vorobej accepts Dispositional Rationality (DR) and IP.[9] But he notes that while it

maximizes to have a policy of retaliation pre-attack, it maximizes to have one of non-retaliation after. So even if, pre-attack, one had the policy of retaliating, one should abandon it after. Moreover, granting it is rational to do whatever is rationally intended, not only is it irrational to act, it is irrational to intend. For by DR it would have become irrational to retaliate post-attack, since the disposition it would then maximize to adopt would be nonretaliatory. By IP, if it is rational to intend, it is rational to act. But since it is not rational to retaliate, by modus tollens, it is not rational to intend it.

Anticipating, Gauthier replied: it is wrong to think that because retaliation is an irrational action--because not maximizing--intending retaliation is an irrational disposition. Rather, because intending is rational--because maximizing--retaliating is rational. But Vorobej sees one man's modus ponens as another's modus tollens. Gauthier's reply would be that one must evaluate the rationality of retaliating by that of the retaliatory disposition because, in a DD, the choice of retaliatory disposition comes temporally first, when it maximizes. But Vorobej could respond that a problem of choice among dispositions rearises before the final choice among actions. (Otherwise, one may not be freely choosing at the point of final choice among actions; one is merely the slave of an earlier choice among dispositions. More below.)

This objection is very compelling. For even if it is maximizing to intend, even in Gauthian rationality that seems separate from whether it will be rational to keep intending post-attack. The conditions making a disposition maximizing might change, and indeed, do so post-attack. Now Gauthier thinks that since this was originally considered in choosing earlier the disposition to retaliate, recalculating later is illicit double counting. But this is wrong. For before, both deterrence and retaliation were possible consequences of a rational choice of disposition. But now, deterrence is impossible, retaliation and its dispreferred harms certain without a change in attitudes. So the conditions determining rational dispositions pre-and post-attack are different.

Granted, a rational DMer acts on maximizing dispositions rather than choosing directly maximizing actions. But if a disposition has not had its designed effect, should a rational DMer conform to it, retaliating? Or should he (if he could) rechoose dispositions, choose a non-retaliatory one post-attack, and not retaliate? If a disposition is rational only if maximizing, then if it ceases to be, one must abandon it for whichever one would now maximize, in this case, one requiring non-retaliation. To act on it would be to not retaliate. So retaliating is irrational even in DR.

Gauthier will say I misconceive the DMer. He would not recalculate the best disposition at each moment. Rather, after noting the consequences of possible dispositions, he would choose a currently maximizing one and then abide by it, even if it failed in its desired effect, so long as he considered this chance in his original choice. It is rational to become someone who stands by old dispositions. Otherwise, one's retaliatory disposition can't deter--no one will believe one would act on it. But this reply has the same problem. Just as the maximizing disposition post-attack is non-retaliatory, so the maximizing kind of agent to be then is one who recalculates/ rechooses dispositions when he has an outdated one which can only be followed now at the cost of performing a pointless non-maximizing action (or, in Gauthier's terms, one no longer expressing a maximizing disposition, since the disposition it expresses has ceased to maximize).

Vorobej's objection returns. Even DR would be known to dictate non-retaliation post-attack. Thus, the DMer knows pre-attack that even he will have reason not to retaliate post-. Thus if it is only rational to intend rational acts, since retaliation is irrational, so is intending it.

Gauthier would reply, of course, that it only maximizes to be a DMer if DMers act on their dispositions in the conditions considered in rationalizing their adoption. It is maximizing and so rational to adopt one only if it will guide post-attack choices. So it must be rational to act on it. But it seems impossible, even by DR, to rationally act in a non-maximizing way post-attack; the non-maximizing effects then of the retaliatory disposition require abandoning it then and so not acting on it. Knowing this, and given IP, one can't rationally intend to retaliate.

4. The Nature of Intentions and the Rationality of Actions

But perhaps we are overlooking something. Assume that if one intends, one will act, perhaps because acting will then somehow be rational, perhaps only because the intention will make one act in a way normally irrational. It maximizes to adopt an act-guaranteeing retaliatory intention. Might intending somehow convey rationality to the action intended? That depends on what an intention is.

(a) Perhaps to intend to do x at time t is just to desire to do something with a certain character, F, at t, and to believe of x that it would have F at t. But our agent prefers not to cause gratuitous harms, and believes retaliating will cause them. So retaliation is irrational. How then can he rationally intend it?[10]

(b) Perhaps an intention is like a promise made to oneself. If one cared to keep promises, one would have reason to retaliate in having promised oneself one would. But our agent does not care to keep promises; only that harms be minimal. (Gauthier himself stipulates this.) He thus has no reason to keep a promise, so the promise could not deter. And knowing its pointlessness, he could not even sincerely make it.[11]

(c) Perhaps to intend is to have a conclusive reason to act. (Gauthier sometimes seems to think of intentions this way.) So perhaps that one has intended gives one a reason sufficient to justify carrying through, especially if one has come to so regard it, has adopted a conception of rationality in which it is decisive (as Gauthier recommends). But post-attack, one has no instrumental justification for regarding one's pre-attack intention as decisive; that can only cause retaliation and gratuitous harms, contrary to the continuing preference that harms be minimal. Thus one has (knowingly or not) conclusive reason post-attack to revert to being someone who does not regard one's original intention to retaliate as decisive. Besides, at best, we have here reason to get oneself to believe that something is a good reason to retaliate. But that needn't mean it is one; it may only justify bad faith, willful irrationality. Moreover, it is odd to think of intentions as being themselves reasons, conclusive or otherwise. Rather, intentions are either what one acquires when one takes oneself to have conclusive reasons for action (but the intention is not itself a reason), or intentions are complexes of reasons for actions--i.e., complexes of beliefs and desires, desires to perform actions with a certain character or consequence, beliefs that the action intended would have that character or

consequence--but intentions are not themselves reasons for action additional to one's beliefs and desires. And as we saw in the reply to proposal (a), above, the beliefs and desires we assume the agent to have in the DD do not appear to justify him in retaliating post-attack.

(d) Perhaps to intend to x at t is to cease to be reflective about what to do at t, so that one acts at t in a non-reflective way; thus one does not revise one's intention at t, but merely acts on it. One's mind is already made up. (This is something which Gauthier himself suggests at one point.) But one will have every reason, post-attack, to be reflective, for that would cause one to revise one's intentions, preventing the gratuitous harms of acting unreflectively on one's prior intention. (One may, in committing to retaliation, have deadened oneself to these considerations. But insofar, surely one would not then, as Gauthier wants, be freely, voluntarily, informedly and rationally choosing post-attack to retaliate. Rather, one would be "choosing" in a non-rational daze, if perhaps in one it was rational to enter.)

(e) Gauthier tends to use terms like "intention," "disposition," "strategy," "plan," "policy," etc., interchangeably. But we might advance by distinguishing between a disposition and an intention. It may be a condition on rationally adopting an intention that the act intended is independently classically rational, i.e., maximizing. But what of dispositions? They may be not so much attitudes towards actions (ones needing to be rational to cause rational actions), as mere habits of choice, reflexes, brain mechanisms empowered to cause behaviors. We might then speak of the rationality of adopting a disposition independently of that of the behaviors it causes, much as we can speak of the rationality of, say, building a machine, independently of its rationality. Here, a disposition is simply a causal determinant of later behavior, an irrevocable reflex, habit, deadman switch. And agents should adopt whichever irrevocable, behavior-determining mechanism would maximize. It will then cause them to behave independently of their preferences (or at least not as straightforward maximizers on their preferences) when it activates.[12]

5. Dispositions as Mechanisms

Perhaps we can now beat Vorobej's objection to the rationality of disposing oneself to retaliate. These things are like doomsday machines, but in people's heads. Vorobej's objection applied here would be that since no one wants a doomsday device to go off, subterranean or subcranial, it is irrational to acquire one. But one does not adopt it because one wants it to go off simpliciter, only to minimize the chance of large harms. Its chance of going off is acceptably low relative to the expectable benefits of having it. Vorobej would be conflating the low utility of it going off with the high expected utility of adopting it. And were dispositions just mechanisms, Gauthier would escape Kavka's objections to the rationality of having a retaliatory disposition from the irrationality of acting on it. For the agent (qua straightforward expressor of a set of preferences concurrent with an action) needn't be able to act (in that sense) on the disposition to rationally have it. It executes itself (or the agent does, but somehow without requiring immediate approval of his then preferences).[13]

This distinguishes what one can directly intend to do from what one can only arrange to have done. It is like a suggestion of Kavka's, that deterrence does not require what appears to be a rationally impossible intention to retaliate, but is attainable by a rational "delegation" of the final "decision."[14] The agent can arrange actions he can't himself perform, hire killers where he can't himself kill, build machines which will inevitably do what he could not bring himself to do directly, induce in himself extinction-resistant, non-reflective dispositions which will by-pass his preferences in guiding his behaviors. Here, the agent qua immediate preference-expressor never has to (classically) choose post-attack to do his own dirty work. His inner doomsday machine uses his body as a retaliatory instrument. He is not, post-attack, (classically) rationally choosing retaliation, of course--merely being guided by the mechanism he chose pre-attack.

Here, the arrangements and the intentions to adopt them have the same effects. The advantages are evident in variations on Kavka's Toxin Puzzle[15]: You are offered $1Million to intend tonight to drink a temporarily nauseating poison tomorrow, but can get the $1Million even if you do not drink. Compare: you are virtually guaranteed there will be no harms if you intend tonight to retaliate if attacked tomorrow, but can get the guarantee even if you do not retaliate. Perhaps you can not rationally intend either to drink or to retaliate, for you disprefer their consequences. But now suppose you are offered $1Million if you will drink the poison right now. Compare: you are guaranteed harms will likely be minimized if you will, right now, press a button on a doomsday machine, guaranteeing retaliation if attacked. Now can you intend to drink the poison, to press the button? Sure, for the balance of reasons favor the action, and so (rationally) should cause an immediate intention to perform it. You will rationally perform it from a rational intention.

You can arrange for what you can't directly intend. For arranging for a retaliation should there be an attack, and retaliating after an attack (independently of any pre-attack commitments), are different actions. So they involve different intentions. (Were they the same, if an agent could do one he could do both. He can only do one so they must be different.) The difference: The rationality of intending depends on the expected utility of the act intended, while that of arranging for an event depends on the expected utility of so arranging. This, of course, depends in part on the utility of the event arranged, but also on its likelihood. The low chance of the harms of retaliation plus the high chance of the prevention of all harms by the arrangement make the arrangement maximizing and so preferable, even if the events conditionally arranged are not.

Now, of course, this does not show the rationality of intending the originally problematic action, but of intending to arrange. It does not prove the rationality of intending where that requires retaliation to be independently (classically) rational. But Gauthier wants retaliation itself to be rational action. Can it so count if caused by these mechanisms? On first glance, no. The agent may choose the mechanism freely and rationally, but it seems implausible to say he thereafter chooses at all, especially freely, never mind rationally. But just as Gauthier argues for a new conception of rationality (or for a consequence of the old), so perhaps we must accept a new conception of freely chosen, voluntary action (or, again, a consequence of the old). Just as retaliation may be rational because intending it maximizes, so perhaps retaliating from a retaliatory

mechanism may be free and rational action, since it would express a disposition freely and rationally chosen.

If so, Gauthier might not just escape its being irrational to form the disposition to retaliate--because irrational (on the old definition) to retaliate--by mooting the rationality of retaliating from the mechanism. Rather, he may conceive it as rational in a different way: Its straightforward (SM) rationality becomes irrelevant. It is only SM rational to adopt the retaliatory disposition, not to retaliate. But it is DR rational to retaliate if that is decided not by its first-order rationality, but by that of the second-order choice of permanent disposition, a choice made pre-attack. Here, it is sufficient for retaliation at t2 = post-attack being a rational action that the adoption of an irrevocable mechanism to cause it is rational at t1 = pre-attack. (Henceforth "permanent" and "irrevocable" mean "permanent/ irrevocable in the envisaged circumstances.") With permanent mechanisms, it is unnecessary for it being rational at t1 to adopt a retaliatory mechanism that that would be rational at t2. Rationality for unrevocable dispositions is different from that for classical intentions. The combination of IP and DR as originally formulated for them is vulnerable to Vorobej; but if we accept the irrevocable mechanism interpretation of the disposition, that formulation is not Gauthier's intent. Instead, he defends retaliation by the principle that an action is rational at t2 just if dictated by the irrevocable disposition maximizing to adopt at t1. This may seem ad hoc, but it seems to yield Gauthier's desired result. If t1 = any time pre-attack, t2 = any time post-, adoption of the disposition seems rational at t1, as does retaliation at t2.

Unfortunately, what makes this a plausible defense of Gauthier is also its undoing. For it is doubtful, first, whether retaliation here is really an action (rational or otherwise) freely and voluntarily chosen and performed at the time of retaliation, even extending Gauthier's rationality into a new conception of action; second, whether retaliation is rational even by this new standard.

6. Irrevocable Dispositions and Action

Gauthier thinks of retaliation from a retaliatory disposition as an action voluntarily, freely, and rationally chosen post-attack. But is it even an action? Maybe not. For (on our current interpretation of Gauthier) it issues from dispositions conceived simply as mechanisms which cause behaviors independently of their bearer's preferences at the time of retaliation (TR). These are just doomsday devices, but in people's heads. And when a nation's doomsday machine self-activates, surely no one has, at that time, chosen and performed a retaliatory action, let alone a rational one. For action chosen and performed now is, normally, informed behavior from a preference concurrent with or just prior to that behavior. (Rational action is behavior one concurrently or just previously instrumentally prefers because it maximizes.[16]) At TR, the disposition makes the agent do things he would rather not do (i.e., cause retaliation's harms). So retaliation is not action.

One can, however, argue that it is an action. For retaliation here would be behavior from a disposition adopted because of instrumental preferences that it be adopted, ones themselves rational because for something--a deterring disposition--maximizing given one's preferences pre-attack. One preferred the disposition, and since one foreknew its

possible conditional effects, surely one conditionally preferred them too. Thus retaliation does issue from one's preferences--the preferences to adopt the disposition--and so should count as action, or at least as part of an action (a rather "fat" one), that of adopting the disposition. Proof: Adoption of the disposition is an action. It consists in determining that retaliation will occur given an attack. So retaliation is or is part of the action of disposition adoption if there is an attack. Details: If retaliation is or is part of the action of the adoption of the disposition, it must satisfy or partly satisfy the same preference as that for the disposition. (Actions are behaviors caused and rationalized by preferences and so are the same just if caused and rationalized by the same ones.) The preference for the specific action of retaliating and that which immediately rationalized adopting the disposition, the preference to adopt it, are the same just if the same events would satisfy them. (That is how preferences get individuated.) Those are the same just if necessarily they exactly overlap spatiotemporally. (That is how events get individuated.)[17] Now the preference to adopt the disposition is satisfied by a change in the bearer's psychological dispositions; the preference to retaliate, by launching some missiles. These may thus seem to be different events, for it seems one could occur without the other. But if one prefers "a mechanism that will retaliate upon attack," the event satisfying the preference to adopt it must be one which includes retaliation by the mechanism given an attack. Otherwise, one would not have adopted "a mechanism that will retaliate upon attack." So the mechanism's retaliation must be in the scope of the preference to adopt it. Thus if adopting it is an action, then the mechanism's retaliation is or is part of it given an attack. (It is a "soft fact" whether adopting the disposition is or has as a part, retaliating. If there is no attack, they are separate events; if there is an attack, the latter is or is part of the former.) Thus both the disposition's adoption and retaliation if there is an attack satisfy the same preferences. So the preference to adopt the retaliatory disposition is the preference to retaliate if attacked. Therefore, retaliation is or is part of the same action as adoption of the retaliatory disposition if there is an attack. Thus, it is made rational and voluntary action by the disposition's adoption being rational and voluntary.

Now this may establish that retaliation is or is a part of, a rational and voluntary action, but maybe not one chosen post-attack. For while nations with rationally chosen doomsday devices, and people with rationally chosen doomsday reflexes are responsible for their devices' effects, this is not because those effects are actions chosen and performed at TR, but because they are parts of actions begun before attack (at TA--Time of initial disposition Adoption).[18]

But this is just refusing to allow Gauthier his new rationality. For if rational choice is choice from a maximizing disposition, the definition of action must change to fit: an action is not a behavior shaped by concurrent preferences, but by a concurrent disposition. So retaliation counts as action on the account which seems to flow from Gauthier's rationality: it is a behavior shaped by a rationally chosen disposition concurrent with or immediately proximal to, the retaliatory behavior. Retaliation might thus seem to be not only an action, but one which, in Gauthier's new sense, is chosen and performed at TR. And we might even have to count it as voluntary (in his new sense), since it issues from a disposition voluntarily chosen at TA.

But there is another sense in which it fails to satisfy even this extended conception of

action, for it doesn't issue from the disposition the agent would volunteer to have, choosing at TR, using DR. To see this and its relevance, we must consider again the rationality of retaliation by Gauthier's standards.

7. Rationality and Irrevocable Dispositions

If Gauthier's is a fully generalizable principle of rationality, it must give consistent results in all applications. Consider now the following times: t1 = pre-attack, t2 = just after attack when one is deciding whether to retaliate, t3 = the time at which one will retaliate, if one is going to. t2 is later than t1, t3 later than t2. Suppose we apply Gauthier's principle that it is rational to do whatever it maximized to be disposed to do, at t1 regarding t3; we get Gauthier's result: it is rational at t1 to adopt, at t1, a disposition to retaliate at t3, and prospectively rational (from the vantage of t1) to retaliate at t3. Suppose now though that, having applied the principle at t1, we try to apply it at t2, just after attack, regarding a final decision of whether to retaliate slightly later, at t3. We will be rational in retaliating at t3 relative to t1's choice. But the principle applied at t2 for t3 would say to adopt, at t2, the disposition re t3 maximizing at t2. That disposition would, of course, be non-retaliatory, since, by t2, an attack has already occurred and there is no longer any point to having a retaliatory disposition: it would only cause more dispreferred harms. Thus the principle now recommends non-retaliation at t3. But one can't conform to the principle applied at t2 re t3, because one is, by hypothesis, irresistibly governed by the disposition adopted at t1 re t3. One may have an excuse for retaliating, but still, barring ad hoc restrictions on substitutions for the time of choice of disposition and the time at which the disposition is directed, the principle that it is rational to act on the disposition it was maximizing to adopt supports "rational" obligations to the conjoint performance of mutually exclusive actions. It is as much a recipe for cognitive dissonance and for behavior which will be in some important sense, unfree and irrational by t3, as it is a recipe for the rational choice of disposition at t1, and a rationalizer of retaliation at t3. It demands that we arrange at t1 for it to be impossible to refrain from retaliating at t3, and then to arrange at t2 not to retaliate at t3. If we do either, we can't do the other. So it is incoherent as a general principle of rationality.

Gauthier tried to show how retaliation from a retaliatory disposition would be free, voluntary, autonomous, and rational action chosen and initiated at TR.[19] To this end, he offered DR (it is rational to do whatever it was rational to be disposed to do) and IP (a behavior inherits rationality from that of the disposition causing it). But these are not really defenses of the claim that retaliation is rational from a disposition to retaliate, only, at best, mere statements of it. And DR as presently understood, seems not even to imply that retaliation would be (unequivocally) rational. For it seems to license reversion, post-attack, to a non-retaliatory disposition (whether this can be done or not--the principle is not sensitive to impossibilities). Moreover, since there are two possible pre-retaliation dispositions--the pre-attack one to retaliate, the post-attack one to refrain--IP gives conflicting recommendations about which should guide one's action. Does retaliating

inherit rationality from that of the retaliatory disposition, or does refraining inherit it from that of the disposition to refrain? In any event, Kavka finds IP indefensible, while Vorobej thinks that it proves the irrationality of adopting the retaliatory disposition. For it finally recommends non-retaliation from the non-retaliatory disposition it is rational to adopt post-attack. Since, by IP, if it is rational to be disposed to do something, it is rational to act on that disposition, and since we have just seen that it is not rational to act on the retaliatory disposition (because it would, rather, be rational to act on the post-attack disposition to refrain from retaliating), by modus tollens, it must not be rational to be disposed to retaliate either. Thus it is irrational to adopt a disposition which would cause retaliation.

We can see now why retaliation is not only not a rational action, but not an action freely and voluntarily chosen post-attack. For then, the maximizing disposition is non-retaliatory, even though, pre-attack, retaliatory. Thus, if one retaliates (from one's pre-attack mechanism), one behaves the wrong way even by what DR standards would later require. DR yields an incoherent standard of the voluntariness of action, and of whether an action has been chosen contemporaneously with the time of the behavior which is to count as action. For an action is both voluntarily and contemporaneously chosen by the standard of issuing from a maximizing (because irrevocable in the circumstances) disposition chosen at t1 = pre-attack re t3 = post-attack, but is neither by the standard of issuing from the maximizing disposition one should choose at t2 = just post-attack re t3 = just pre-choice of whether to retaliate. This is so precisely because of Gauthier's taking the rationality of retaliation to be decided by whether it issues from a maximizing disposition.

Perhaps one's having preferred the disposition makes retaliation at least part of an action initiated by or consisting in part in, the disposition's adoption. But that does not make it an action both chosen and performed post-attack; only, at best, one initiated long before under different conditions, one it would now (post-attack) be irrational to choose to perform, and which one rationally would now have to truncate if one could. In that sense, it is not now voluntary. It finds one doing something one ought now to choose not to do, and which one would now choose not to do, if one could.[20]

8. The Gauthier Replies

Gauthier would make three replies. First, we misinterpret DR. The DR agent chooses from the disposition the having of which maximizes, but not necessarily the one maximizing at TR. He can rationally choose from the one it maximized to adopt when the conditions under which a final choice among actions would be made were merely prospective. Even though, post-attack, the rational disposition is non-retaliatory, it can remain rational to choose from the one it was rational to adopt pre-attack. It is rational to act on a rational plan about how to act.

But there two relevant occasions when the final choice among actions is prospective, namely, before attack, and just after, but just before retaliation. One can thus plan anew. And while it was rational pre-attack to dispose oneself to retaliate when the final choice among actions was still in the future, it is rational post-attack to dispose oneself to non-

retaliation given that there has been an attack, where the final choice among actions is still ahead. There seem then to be two possible dispositions prior to the final choice among actions that compete for being the correct determinant. And I can think of no non-ad hoc way of arguing that the non-retaliatory one should not govern one's behavior post-attack. For it maximizes at TR, is temporally closest to whichever action is to be chosen, and is in every other way the one which would normally be relevant as an intention serving as the rational determinant of an action. (This makes it impossible to deter with the retaliatory disposition pre-attack if one could shed it post-; but as we saw, being moved by this is just wishful thinking.)

Indeed, surely one decides whether a disposition continues to be rational to have and act on by whether it maximizes to (continue to) have it. That, after all, was the basis for choosing Gauthier's preferred disposition at TA. It ceased to maximize to be a SMer bereft of a retaliatory disposition at TA, which was why one became a DMer with a retaliatory disposition. How can one satisfy the rational obligation to always choose, and to choose from, a maximizing disposition, while choosing from one that is non-maximizing post-attack? And doesn't it cease to maximize to be a CMer with a retaliatory disposition at TR?

Nonetheless, mightn't it have been rational to make an irrevocable commitment to a retaliatory disposition (irrevocable as a matter of causal fact, that is; for it is not clear how one could render it rationally irrevocable)?; otherwise, it would not deter. Whether one can rationally do this in Gauthier's scheme, however, is still doubtful, because it may result in a non-maximizing behavior which one could, post-attack, neither directly intend (by SM) nor retain a rational disposition to perform (even by DR, applied post-attack) given one's preferences. As we saw earlier, even in DR, by the time of possible retaliation, one is rationally obliged to be disposed not to retaliate; it is therefore irrational to retaliate; since, if it is rational to be disposed to retaliate it must be rational to retaliate, and since it is not rational to retaliate, it cannot be rational to be disposed to retaliate either. In any case, whether it is rational pre-attack to adopt a disposition one cannot revoke post-attack is separate from whether it would be rational to revoke it if one were given that option post-attack. And nothing in either standard rationality or in Gauthier's alternative suggests that that is not the relevant counter-factual in deciding the rationality of anything able to count as a genuine choice post-attack (however much Gauthier would like it to be otherwise.) It seems it would be rational to revoke it, for a disposition is made rational, on Gauthier's own account, by its being maximizing, and the retaliatory one isn't, post-attack.

Gauthier's second response: we beg the question against him in evaluating the rationality of both action and disposition by whether each now serves the agent's current preferences. Rather, to be rational is to evaluate choices (on their rationality, and, presumably, their freedom, voluntariness, status as genuine actions, and status as now-chosen actions) not by one's current preferences, but by one's rationally chosen dispositions. Indeed, he may think that further possible dispositions should be evaluated by whether their adoption expresses the original one. And of course, the retaliatory one would not be expressed in revoking it here.

But surely the only possible argument for its being rational to evaluate an action or

disposition by its concordance with an earlier-adopted disposition, is that the earlier one is rational, the one it is currently rational to have guide action. Gauthier's own standard for the rationality of a disposition is whether having it maximizes. And while the retaliatory disposition was a rational (because maximizing) one to have adopted pre-attack, it is not rational to have it post-(for it no longer maximizes). So by Gauthier's standard for whether a disposition and choice from one is rational--vis., whether the disposition is maximizing, the choice, one expressing a maximizing disposition--the retaliatory one is not rational post-attack; nor then is retaliation.

True, it would maximize to adopt the retaliatory disposition if it would guide post-attack behavior. And it would confirm the rationality of deterrence and of action on deterrence policy were it then rational to use that disposition in the final decision of whether to retaliate. But surely that it would be pre-attack maximizing were that the rationally correct disposition to use does not by itself make it correct. Indeed, if Gauthier is not to beg the question, he must offer a standard of the rationality of a disposition other than that it would be maximizing if it were rational. And he has: a disposition is rational if maximizing, a choice of action, if it expresses the maximizing disposition. But by that standard, it is neither rational post-attack to have a retaliatory disposition, nor to choose from it.

Finally, Gauthier will claim that if the agent is to rationally deter by threat of rational retaliation, he must adopt a disposition which makes it rational to retaliate; otherwise, it would not maximize to adopt it (for an agent who will later choose rationally). But this does not establish that there is a disposition whose adoption rationalizes retaliation, only that it would maximize were there one.

Gauthier would find these objections maddening. For the whole point of adopting a disposition is to secure a later behavior it is earlier advantageous to have secured; yet it would seem that, later, it is the wrong disposition to have guide one's behaviors.

One might conclude that it is impossible to rationally and efficaciously plan now to rationally perform inutile future actions in paradoxical choice situations. For while there is advantage to adoption of the plan, there is no advantage to action upon it. But I think this only shows that rationally committing to such plans must have a different character than Gauthier thought. It must consist in one's preferences rationally changing so as to make compliance with the commitment/plan classically rational.

9. Rational Plans in Paradoxical Choice Situations; The Rational Preference to Retaliate

Gauthier's conclusions can be saved from Vorobej's and Kavka's objections, if we interpret the retaliatory disposition as a set of preference-functions on which it maximizes to retaliate.[21] Here, one again faces a second-order choice problem. But one chooses among preference-functions over outcomes or actions (i.e., among orderings of outcomes and actions from most to least preferred). One chooses from one's original preferences for outcomes of choice and one's beliefs about the strategic effects of having new preferences. One should adopt whichever ones would deter, i.e., ones maximization on which requires retaliating. E.g., one becomes such that one prefers that harms be minimal unless one is attacked, in which case, one prefers to retaliate. Enemies would then be

deterred, seeing one would then find it rational (because maximizing given one's new values) to retaliate. Since one has such preference-functions, one must retaliate if attacked, as a rational SMer.

Note that the new preference-orderings serve all the roles Gauthier wished for dispositions: it is deterring and so maximizing to adopt them; it is rational to retaliate from them, since they are preferences to retaliate; their very adoption conveys rationality to retaliating (vindicating IP for this type of choice); rationality proves to require that choices be made on the basis of a second-order choice among determinants of further choice. But the principle which justifies adopting these bases of further choice--vis., that one should always maximize on whichever preferences one has at the time of choice--is coherent in these applications: it maximizes to adopt new preferences pre-attack, maximizes (given that one has new preferences) to keep and act on them post-. We can now solve the intention problem, which was: To rationally intend to do x, one needs conclusive reason to do x; but since one disprefers retaliation's harms, one has no reason to retaliate, and so can't intend it. So one would not retaliate given the choice. Thus, one can not deter by intending it. Our solution: It would best serve the preference that harms be minimal to acquire a modified set of preferences in which one prefers that harms be minimal only provided attacks are met with retaliation. So preferring, one now has a reason to retaliate, for it now maximizes on one's new preferences. And one's intention will now deter, for were one rational, one would act on it. So it is now rational to intend to retaliate.

Now Kavka and David Lewis considered a similar proposal.[22] But both thought it would not be rational to acquire a preference to retaliate and to act on it, because just that you have such a preference does not mean retaliating is all things considered rational given the totality of your preferences. Your new preference might not accord with the rest; action on it might not maximize given all of them. In that case, you should refrain from acting on it post-attack, and indeed, should give it up. But I say one must not just acquire a preference for retaliation, but must revise one's total preference set into one in which retaliation is preferred over everything else if there has been an attack, and in which the minimization of harms is otherwise always preferred. Since it is rational to act as one most prefers, and since, post-revision, one most prefers to retaliate given an attack, that is rational.

10. Conclusion

How do we come down on our original four problems? Well, one cannot rationally intend or rationally perform a non-maximizing action, but it is sometimes rational to so change one's preferences as to convert a non-maximizing action into a maximizing one, upon which one can both rationally intend and rationally perform it. The rationality of intending and acting are linked: one can only rationally intend a maximizing and so rational action. But whether an action is rationally intendable and performable depends on the prior question of whether one's current preferences regarding it are rational; and they are
not if it maximizes to revise them so that one can adopt an intention advantageous by

their original standard. This position is odd in its own ways, doubtless now troubling the reader. Admittedly, there are puzzles, though I think they can all be solved. Most of these I will have to leave for another occasion[23], though I will speak briefly of some of them in the next section. But I hope I've already said enough to show that this position has what Gauthier desired in his solution, whatever its own vices.

More importantly, the problems Gauthier faced are instructive for thinking about rational intentions and plans in general. They show that the fields of action theory and decision and game theory need radical revamping. For all these problems arise from two assumptions common in those fields, first, that a choice is practically rational if it advances one's preferences, second, that instrumental rationality involves only how to choose given one's preferences. This, of course, engenders paradox, for it can advance one's preferences to commit to the performance of actions that do not advance one's preferences. But the paradoxes only arise through a narrow-mindedness about the power of instrumental rationality to adjudicate the rationality of preferences themselves. Indeed, the first dogma falsifies the second: If it is rational to do whatever will advance one's preferences, since it will sometimes advance one's preferences to have different preferences, it is false that a rational choice is always a choice given one's current preferences. In fact, a choice is only rational if it advances preferences it is rational to have; and it is irrational to keep preferring something if revising one's preferences would advance their own satisfaction. This explains how it is possible to advantage oneself by a rational commitment to perform an initially non-advantageous act: it is rational to change what one prefers where it advances what one prefers not to prefer it, namely, where it most advances one's preferences to so prefer as to be able to rationally perform an action that does not advance one's original preferences.

11. Prolegomenon

I acknowledged earlier that my proposal might seem to give rise to its own puzzles and paradoxes. I cannot take up all of these possible worries here, but perhaps I should say a word about one or two of them. First, it might be thought that the preference to retaliate if attacked is mad. Well, perhaps it is, in some moral sense. But my only concerns here are with whether it is instrumentally rational to acquire it given that one begins with preferences which it advances to supplant with this odd preference, and whether an agent in the situations here considered can be said to choose rationally if he does not first so revise his preferences. If pressed, however, I would also argue that the retaliatory preference is not morally mad or monstrous in the circumstances, that is, considering the context of its acquisition. But that would take at least another paper to show.[24]

Secondly, it might be thought that, since the maximizing conception of rationality says it is rational to choose so as to advance one's current preferences, and since I am recommending that agents sometimes change their preferences as a means of advancing their original preferences, my proposal deprives the maximization standard of its frame of reference in present preferences. I do not believe that my proposal has this effect, however. Rather, it expressly preserves that frame of reference. At any time of choice, one is to choose from whatever preferences one then has. But some circumstances in

effect present one with a problem of choice among preferences given one's present preferences. Such, I claim, is the case in the DD. Here, before choosing among actions, one must first choose among possible preferences. The reason: to choose here immediately among intended actions without first choosing among preferences, is non-maximizing; it maximizes to choose first among preferences. Since one's enemies can be deterred only if they see that one would find it rational to retaliate, and since one would only find it rational to retaliate if one preferred to retaliate given an attack, one must choose to acquire as one's strongest preference, the preference to retaliate if attacked. One's initial preference that harms be minimal thus justifies as a maximizing choice, choosing to have new preferences. Thereafter, if one is attacked, since one then has new preferences, and since it is always rational to maximize on whatever preferences one has when facing a choice, one must, to maximize on the new preference to retaliate if attacked, now choose to retaliate. In this second choice, one does not face a choice among preferences, since one now has no reason to change one's preferences again. (One no longer prefers that harms be minimal simpliciter, so that old preference can give one no reason to revert to one's original values.) Given that one now prefers that retaliation occur, the only thing that could make it rational to change that preference, is if so changing it were necessary to causing retaliation. E.g., suppose the attackee were accosted by someone paternalistically concerned for his character. This paternalist will not let the attackee retaliate, but will retaliate for him provided the attackee comes to prefer that there be no further harms. The attackee would then find it maximizing to acquire pacifist preferences as a means to securing retaliation. But the attackee does not in fact have to deal with the paternalist in the classic Deterrence Dilemma. Rather, post-attack, he faces a choice among retaliating or refraining from retaliating. And since he prefers to retaliate if attacked, and since he has now been attacked, it is now maximizing for him to retaliate. So: one maximizes, prior to a possible attack, by acquiring a dominant preference to retaliate if attacked; and, after attack, one maximizes on that new preference in retaliating. Throughout, one always maximizes on the preferences one has at the time of choice.[25]

Finally, it may seem odd that one could satisfy a preference by changing it. If I no longer have a preference for x, how can the obtaining of x yield me any satisfaction? To give an example rather more stark than that provided by the DD, consider the Gift Dilemma: I will give you something, x, which you now want very badly to have, just if you come to want very badly not to have it. Now it may seem that if you come to disprefer x, me then giving you x will not satisfy you; and if a rational agent always aims at his own satisfaction, it may then seem that it is not rational to change your preferences.

But this worry implies a misunderstanding, I think, of the nature of preference satisfaction and of the way in which it is relevant to the rationality of choices. First, let us distinguish between satisfaction and utility. One's preference for x is satisfied simply by the obtaining of x. Utility is what one gets when one of one's preferences is satisfied at the same time as one has that preference. So you get utility from the satisfaction of a preference for x just in case x comes to obtain while you still prefer that it obtain. But rational agents must maximize not their individual utility, but their individual expected utility. That is, they must choose so as to make as high as possible the product of the

probability and preferability by the measure of their current preferences, the outcomes of their choices. Plainly it is possible for circumstances to be such that something one prefers to obtain can only be caused to obtain by one's ceasing to prefer it. In the Deterrence Dilemma, I can only cause the obtaining of conditions with no harms in them by ceasing to prefer always to minimize harms. In the Gift Dilemma, I can only cause the obtaining of the condition of me owning x by ceasing to prefer owning x. In both dilemmas, by changing my preferences, I cause them to be satisfied. Now it is true that, in the case of the Gift Dilemma, the satisfaction of my original preference will not give me any utility, for utility is the obtaining of a preferred condition while it is preferred, and to cause the obtaining of the originally preferred condition, I had to cease preferring it by the time of its obtaining. But this is no objection to the rationality of me changing my preferences, for the aim of rational agents is to try to cause the satisfaction of their preferences, not the utility of satisfying them. And in changing my preferences in the Gift Dilemma, I do so cause their satisfaction, for I cause to be likely the obtaining of the conditions I originally preferred at the time of choosing among preferences. Matters would have been different had I had not just the preference to own x, but also the preference to enjoy it while owning it. One can only enjoy owning x if one still prefers to own it at the time one comes to own it. So it would not have been maximizing for me to revise my preferences if they had consisted in both a preference to own x, and a preference to enjoy the owning of x. But not all preference functions contain both a preference for something and the preference to enjoy the obtaining of that something. E.g., in preferring that one's family inherit one's fortune when one dies, one is not at the same time preferring to be around to enjoy their financial security when one dies; one knows perfectly well that one won't be around to enjoy anything when one is dead. But that doesn't mean it is somehow irrational to prefer that one's family inherit one's fortune, nor irrational to take steps to secure that condition, e.g., to make out a will. Indeed, I offer the case of inheritance as a counter-example to any philosophers who think that the aim of all rational choice is utility--the enjoyment of conditions--and not mere satisfaction--the obtaining of conditions that one currently prefers to have obtain at some time. (The only cases in which one must aim at utility are those in which one's preferences are, in effect, preferences <u>for utility</u>, as I have defined it. And even in those cases, one tries to cause utility not because doing so is integral to the structure of rationality, but because causing oneself to have utility is, in those cases, necessary to causing the satisfaction of one's preferences; for in those cases, they are preferences <u>for utility</u>.) The only way I can rationally prefer that events obtain after my death, and the only way I can rationally choose in light of such preferences, is if it is not a necessary condition of the satisfaction of such preferences, that I will still (be around to) have them at the time of their satisfaction; that is, if it is not necessary to the satisfaction of every preference that the condition it is a preference for obtain concurrently with the obtaining of that preference.[26]

Notes

1. For helpful discussion and/or correspondence, I am grateful to Neera Badhwar, David Braybrooke, Robert Bright, Douglas Butler, Peter Danielson, Gregory Kavka, Carl Matheson, Robert Martin, Victoria McGeer, Geoffrey Sayre-McCord, Howard Sobel, Terry Tomkow, Kadri Vihvelin, Michael Webster, Sheldon Wein, an anonymous referee for this Journal, and especially Julia Colterjohn and Richmond Campbell. My thanks as well to the audience at the 1988 meetings of the Canadian Philosophical Association to which part of an earlier version was presented, especially to the commentator, David Zimmerman. I thank Dalhousie University for the Killam Post-Doctoral Fellowship which funded early work on this project.

2. David Gauthier, "Deterrence, Maximization, and Rationality," Ethics 94 (1984), pp. 479-480.

3. Gregory Kavka, "Some Paradoxes of Deterrence," in John Perry and Michael Bratman, Michael, eds., Introduction to Philosophy: Classical and Contemporary Readings (New York and Oxford: Oxford University Press, 1986), pp. 516-526. Originally published in The Journal of Philosophy, 75 (1978), pp. 285-302.

4. Kavka, "Some Paradoxes," and Gregory Kavka, "Responses to the Paradox of Deterrence," in Douglas Maclean, ed., The Security Gamble: Deterrence Dilemmas in the Nuclear Age (Totowa, N.J.: Rowan and Allenheld, 1984), pp. 155-159.

5. See Gauthier, "Deterrence," pp. 479-480, 482-483, 486-488; also, David Gauthier, Morals By Agreement (Oxford: Clarendon Press, 1986), Chs. I, V, VI, and IX.

6. The following arguments are from Gauthier, "Deterrence," pp. 479-480, 482-483, 486-489, and from David Gauthier, "Afterthoughts," in Maclean, The Security Gamble, pp. 159-161.

7. Mark Vorobej, "Gauthier on Deterrence," Dialogue: Canadian Philosophical Review, XXV (1986), pp. 471-476.

8. E.g., see J. Howard Sobel, "Maximizing, Optimizing, and Prospering," Dialogue: Canadian Philosophical Review, XXVII (1989), pp. 233-262.

9. Vorobej, "Gauthier."

10. I will argue that it is rational to change one's preferences so that one could then intend to retaliate, having reason to keep to the intention in one's new preferences. But that is not Gauthier's solution for it alters one's over-all preference-functions, which he thinks can remain constant.

11. It might be rational to come to prefer the keeping of promises, then to promise. But that is not Gauthier's solution, for he holds preferences constant.

12. It is sometimes tempting to read Gauthier as suggesting something like this for the Prisoner's Dilemma. E.g., see Gauthier, Morals By Agreement, pp. 2, 3, 79, 158, 168-9, 188.

13. The agent may object to adopting such devices to secure deterrence. But then, as David Lewis would point out, he is not in a DD; his predicament is irrelevant to our case. See his, "Devil's Bargains and the Real World," in Maclean, The Security Gamble, p. 142.

14. Kavka, "Some Paradoxes," p. 521.

15. Gregory Kavka, "The Toxin Puzzle," Analysis 43 (1983), pp. 33-36.

16. Some actions may be preferred for their own sakes, but retaliation is here initially dispreferred.

17. I adopt Myles Brand on individuating intentions, actions and events. See his, Intending and Acting: Toward a Naturalized Action Theory, (A Bradford Book, The MIT Press: Cambridge, Massachusetts, 1984).

18. For more on this, see my, "Libertarian Agency and Rational Morality: Action-Theoretic Objections to Gauthier's Dispositional Solution of the Compliance Problem," The Southern Journal of Philosophy, XXVI (1988), pp. 499-525.

19. I argue, in my "Libertarian Agency," that this is impossible for such dispositions in general, by examining how similar ones may figure in Gauthier's attempted solution to the problem of guaranteeing compliance with agreements to co-operate in Prisoner's Dilemmas.

20. An analogy: I resolve to kill you with slow-acting poison in your soup. You drink it; I have a change of heart. Too late, there is no antidote; you die with me holding your hand, weeping apologies. Now, I freely, voluntarily, and rationally killed you. But I did not choose, after you drank, to kill you. And killing you is not a free, rational, and voluntary action which I chose to perform after you drank. (Thanks to Robert Martin for this point.)

21. For an attempt to show that, by his own conception of practical rationality, his theory of the rationality of co-operation only works if the CM disposition is a revised preference set with an over-riding preference for co-operation with those with similar preferences, see my "Two Gauthiers?," Dialogue: Canadian Philosophical Review XXVIII (1989), pp. 43-61, and my "Co-operative Solutions to the Prisoner's Dilemma," Philosophical Studies 64 (1991), pp. 309-321.

22. See Lewis, "Devil's Bargains," pp. 153-154, and a comment in the notes to Kavka, "Responses."

23. See notes 24-26, below.

24. For more on this, and for an attempt to make the idea of practically motivated, rational revisions in one's preferences seem less counter-intuitive, see my, "Preference Revision and the Paradoxes of Instrumental Rationality," forthcoming, Canadian Journal of Philosophy, and my "Kavka Revisited: Some Paradoxes of Deterrence Dissolved," unpublished manuscript, Dalhousie University, 1990

25. For more on how my proposal preserves the frame of reference of rational choices in present preferences, see my "Preference's Progress: Rational Self-Alteration and the Rationality of Morality," Dialogue: Canadian Philosophical Review, 30 (1991) 3-32. In that paper, I also apply the proposal to the analysis of morality as a solution to the Prisoners' Dilemma, and relate the proposal to the conative structure of conclusive reasons for action.

26. For more on the apparent paradoxes in the idea of revising preferences to satisfy them, and for an attack on other conceptions of practical rationality, see my "Persons and the Satisfaction of Preferences: Problems in the Rational Kinematics of Values," unpublished manuscript, Dalhousie University, 1991. My thanks to the reader for this Journal for raising some of the issues discussed in the last section.