RATIONAL COOPERATION, IRRATIONAL RETALIATION

BY

JOSEPH MINTOFF

David Gauthier claims it can be rational to perform a non-expected-value maximizing cooperative act.[1] We can construct a simple version of the argument as follows: (1) there are situations—which I will call Special Cooperation Situations—in which it is rational to adopt an intention to perform a non-expected-value maximizing cooperative action; but, (2) if it is rational to adopt an intention to do something, then it is rational to do that thing; therefore, (3) there are situations in which it is rational to perform a non-expected-value maximizing cooperative action. I will call this the **Cooperation Argument**.

One type of objection to this argument focuses on the supposed rationality of adopting the intention to perform a non-expected-value maximizing cooperative action. Some object to the possibility of non-maximizing actions, claiming that an action can be intentional, or free, only if it is expected-value maximizing.[2] Others object to the possibility of adopting an intention to perform such an action, claiming that one can adopt an intention to perform some action only if the action is expected-value maximizing.[3] Further, some point out that such an intention is rational only if intentions in general are transparent or translucent, and that in reality they are not so.[4] In this paper, though, I will not be concerned with this type of objection.

Another type of objection focuses on the move from the rationality of adopting the cooperative intention, to the rationality of the cooperative action itself. Gregory Kavka, for example, makes this point.[5] We can construct a simple version of the objection as follows: (1) there are situations—which I will call Special Deterrence Situations—in which it is rational to adopt an intention to perform an irrational retaliatory action; (2) Special Deterrence Situations and Special Cooperation Situations are relevantly similar; and so (3) in a Special Cooperation Situation, it is rational to adopt an intention to cooperate, but *irrational* actually to cooperate. I will call this the **Deterrence Objection**. In this paper, I examine one way in which the Cooperation Argument may be reformulated to circumvent this particular objection.

Pacific Philosophical Quarterly 74 (1993) 362–380 0279–0750/93/0400–0000
© 1993 University of Southern California. Published by
Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and
238 Main Street, Cambridge, MA 02142, USA.

The Deterrence Objection is, in my view, a particularly forceful objection to the Cooperation Argument. One may indeed be convinced that there are realistic situations in which one can, and it is rational to, adopt an intention to perform a non-maximizing action, and yet not be convinced that the intended action inherits that rationality. This seems particularly clear in Special Deterrence Situations. As we will see when we examine them in more detail below, many people would be convinced that it is irrational to retaliate in Special Deterrence Situations, since such retaliation results in no benefit, but only horrendous destruction. And many would also be convinced that Special Cooperation Situations and Special Deterrence Situations are relevantly similar—each concern rationally adopted intentions to perform non-maximizing actions. What I have called the Cooperation Argument is a crucial move in Gauthier's attempt to argue that it could be rational to act morally, and, for this reason, an examination of the ways in which it may be reformulated to avoid this objection is worthwhile.

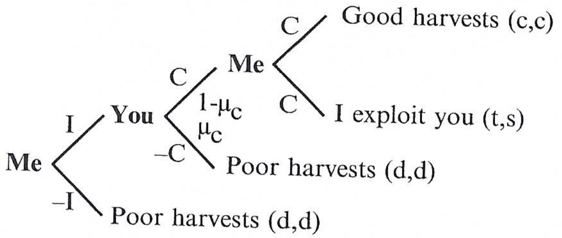
The paper has three sections. In the first, I introduce the Cooperation Argument and the Deterrence Objection in more detail. In the second, I introduce one possible reformulation of the Cooperation Argument, by replacing its second premise with a principle connecting rationally adopted intentions, rational action, and rational reconsideration, and a specific theory of rational reconsideration. In the final section, I argue that this reformulated Cooperation Argument is not susceptible to any form of the Deterrence Objection. I conclude that the Deterrence Objection may indeed be circumvented if proper attention is paid to the role of rational reconsideration.

I. *Cooperation and Deterrence*

In order to see how the Cooperation Argument may be reformulated, we need to examine it, and the Deterrence Objection, in more detail. In the process I provide a fuller account of what I have called Special Cooperation Situations and Special Deterrence Situations.

THE COOPERATION ARGUMENT

You and I have adjacent farms.[6] Although neighbors, and not hostile, we are also not friends, so that neither gets satisfaction from assisting the other. Next week, my crops will be ready for harvesting; a fortnight hence, your crops will be ready. We recognize that if we harvest our crops together each does better than if each harvests alone. The harvest in, I am retiring, selling my farm, and moving to a retirement village, where I am unlikely to encounter you or any other members of our community. I'd like to promise to reciprocate cooperation—that is, I'll help you if you help me first—but you can tell whether or not I really intend to do so, and my intending to do so is necessary and very likely sufficient for you to cooperate. We may represent the situation we face diagrammatically as follows:



$t > c > d, \mu_C < 1$

Figure 1[7]

The expected-values I assign to the possible outcomes are indicated by the left-hand constant, and are: my exploiting you, by accepting your help with my harvest, but not helping you with yours (= *t*, Temptation payoff); each helping the other with their harvest, thus resulting in good harvests for each (= *c*, Cooperation payoff); and neither helping the other with their harvest, thus resulting in poor harvests for each (= *d*, Defection payoff). Clearly, $t > c > d$, since I value most the outcome of my exploiting you, second a good harvest, and third a poor harvest. The values you assign to the possible outcomes are indicated by the right-hand constant, and, in addition, include: my exploiting you (= *s*, Sucker payoff). Clearly, $c > d > s$, since you value most a good harvest, next a poor harvest, and last my exploiting you.

Two things follow in such a situation. First, that it maximizes expected-value for me to adopt[8] what I call the reciprocation intention: that is, the intention that if you cooperate first then I will cooperate (and if you do not, then not). It maximizes expected-value provided that the probability, μ_C, that you do not cooperate, given that I have adopted this intention, is less that one.[9] In other words, if there is some chance that you will be convinced by the presence of this intention, then I have nothing to lose, and possibly something to gain, by adopting it. Second, it follows that, even if you were to cooperate, it still maximizes expected-value for me not to cooperate in return. The day before my retirement, helping you is a dead loss to me.

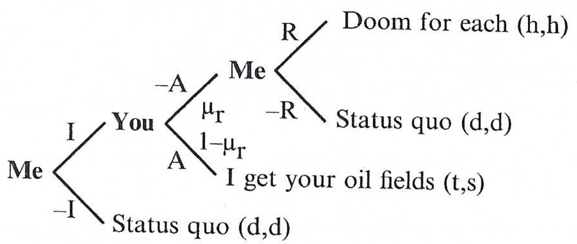
We find ourselves in what I call a **Special Cooperation Situation** (or an SCS for short).[10] More generally, an agent is in an SCS when he reasonably and correctly believes that the following conditions hold. First, he must adopt the reciprocation intention if the other is to cooperate. Second, the adoption of such an intention would very likely induce the other to cooperate. Third, the amounts of benefit involved are very large, and are such that a consequentalist calculation would substantially favor adopting the intention. Finally, it would have the best outcome for him not to cooperate, if the other were to cooperate first. In short, it maximizes expected-value to adopt the intention, since it is necessary and very likely sufficient to induce the other to cooperate, but maximizes expected-value not to carry it out.

The Cooperation Argument is now simple to state more precisely. (1) In an SCS, it is rational to adopt the reciprocation intention, which is an intention to perform a non-maximizing cooperative action. It is rational to adopt this intention because it maximizes expected-value to do so. However, (2) "[i]f it is rational for me to adopt an intention to do *x* in circumstances *c*, and if *c* come about, and if nothing relevant to the adoption of the intention is changed save what must be changed with the coming about of *c* ... , then it is rational for me to carry out *x*."[11] Since (we may suppose) nothing relevant to the adoption of the intention has changed save what must have changed with the coming about of your cooperation, it follows that (3) there are some situations in which it is rational to perform a non-maximizing cooperative action.

THE DETERRENCE OBJECTION

Some are not happy with this argument, and in particular are not happy with its move from the rationality of adopting an intention, to the rationality of performing the intended action. In response they introduce so-called Special Deterrence Situations to argue that this move is invalid.

You and I are the despotic leaders of adjacent nations.[12] We are certainly not friends, especially since we both know I covet the oil fields just the other side of my border with you. If you do not willingly give them to me, then my only option would be to start a devastating war which would benefit neither of us. I'd like to get you to cede the oil fields to me, so I am thinking about adopting a deterrent[13] intention: to retaliate with war if you do not acquiesce. I know you can tell whether or not I really intend to do so, and my intending to do so is necessary and very likely sufficient for you to acquiesce. We may represent the situation we face diagrammatically as follows:



$t > d > h$, $\mu_r < (t–d)/(t–h)$

Figure 2[14]

The expected-values I assign to the possible outcomes are: my getting your oil fields, by having you acquiesce to my threat (= t, Temptation payoff); the status quo, where I do not get your oil fields, but we avoid all-out war (= d, Defection payoff); and all-out war, if I do retaliate after your non-acquiescence (= h, Holocaust payoff). Clearly, $t > d > h$, since I value most the outcome of getting your oil fields, second the status quo, and a distant third the destruction of my nation. The values you assign to the possible outcomes, in addition, include: your ceding your oil fields to me (= s, Sucker payoff). Clearly, $d > s > h$, since you value most the status quo, next ceding your oil fields, and a distant third the destruction of your nation.

Two things follow in such a situation. First, that it maximizes expected-value for me to adopt what I call the deterrent intention: that is, the intention that if you do not acquiesce then I will retaliate (and if you do, then I will not). It maximizes expected-value provided that the probability, μ_r, that you do not acquiesce, given that I have adopted this intention, is less that $(t–d)/(t–h)$.[15] In other words, if there is a sufficiently small chance that you will not be convinced by the presence of this intention, then I can expect to do best by adopting the intention. Second, it follows that, even if you were not to acquiesce, it still maximizes expected-value for me not to retaliate in return. The day after you have snubbed my threat, plunging both our nations into a destructive war is simply a dead loss to me.

We find ourselves in what I call a **Special Deterrence Situation** (or an SDS for short).[16] More generally, an agent is in an SDS when he reasonably and correctly believes that the following conditions hold. First, he must adopt the deterrent intention if the other is to acquiesce. Second, the adoption of such an intention would very likely induce the other to acquiesce. Third, the amounts of benefit involved are very large, and are such that a consequentalist calculation would substantially favor adopting the intention. Finally, it would have the best outcome for him not to retaliate, if the other did not acquiesce. In short, it maximizes expected-value to adopt the intention, since it is necessary and very likely sufficient to induce the other to acquiesce, but maximizes expected-value not to carry it out.

The Deterrence Objection is now simple to state more precisely, and it is easy to see why it is so forceful. (1) In an SDS, it is rational to adopt a deterrent intention to perform an irrational retaliatory action. It is rational to adopt this intention because it maximizes expected-value to do so; it is irrational to retaliate because, similarly, it maximizes expected-value *not* to do so. Consequences may not matter all the time, but they certainly do when the stakes are high—and in an SDS the destruction of my nation is at stake. (2) SDSs and SCSs are relevantly similar. This should now be abundantly clear from the way we have introduced them—simply compare Figures 1 and 2. Hence (3) in a SCS, it is rational to adopt an intention to cooperate, but *irrational* actually to cooperate.

II. Cooperation, Deterrence and Rational Reconsideration

The Deterrence Objection is a strong one. In this second section, I introduce a principle connecting rationally adopted intentions, rational non-reconsideration, and rational action, and a specific theory of rational reconsideration, which together can be used to reformulate the Cooperation Argument to avoid this objection.

A PRINCIPLE OF RATIONAL INTENTION, RECONSIDERATION, AND ACTION

The issue I am concerned with in this paper is the move from the rationality of adopting an intention to the rationality of acting on that intention. The principle of primary focus in this dispute is thus the claim that if it is rational for me to adopt an intention to do x in circumstances c, and if c come about, and if nothing relevant to the adoption of the intention is changed save what must be changed with the coming about of c, then it is rational for me to carry out x. Call this **Gauthier's Claim**. The Cooperation Argument uses Gauthier's Claim as its second premise; the first premise (and conclusion) of the Deterrence Objection entails that Gauthier's Claim is false.

Yet there are other, simpler, objections to Gauthier's Claim. First, the term 'rational' is ambiguous between 'rationally obligatory' and 'rationally permitted', but Gauthier's Claim does not make clear which is intended. Second, even if it is rational to adopt the deterrent intention, perhaps (irrationally) I have not adopted it. And surely it is not rational to retaliate when one has not even bothered to adopt the deterrent intention, and, as a result, the other has not been deterred. Third, even if circumstance *c* comes about, perhaps (lacking any evidence) I am rationally permitted not to believe it, or (having evidence) I do not believe it has come about. And surely it is not rational to retaliate when, rationally or irrationally, one does not believe the other has failed to acquiesce. Finally, even if nothing relevant to the adoption of the intention is changed save what must be changed with the coming about of circumstance *c*, perhaps (rationally or irrationally) I have lost the deterrent intention. And surely, in such a case, it is not rational to retaliate. For all these minor reasons, then, one might independently be inclined to reject Gauthier's Claim, and with it the Cooperation Argument.

The initial step in reformulating the Cooperation Argument is to replace Gauthier's Claim with a weaker principle taking into account all of these minor objections (the italicized expressions indicate the differences):

(G*) if *I rationally ought* to adopt an intention to do *x* in circumstances *c, and I do adopt it*, and if *c* come about, *I rationally ought to believe c has come about, and I do believe it*, and if nothing relevant to the adoption of the intention is changed save what must be changed with the coming about of *c*, *and I still have this intention*, then *I am rationally permitted* to carry out *x*.

All of these changes make principle (G*) weaker than Gauthier's Claim: to replace 'it is rational for me' by 'I rationally ought' in the antecedent is to make the antecedent no weaker (and perhaps stronger); to add the clauses 'and I do adopt it', 'I rationally ought to believe c has come about, and I do believe it' and 'and I still have this intention' is to make the antecedent stronger; to replace 'it is rational for me' by 'I am rationally permitted' is to make the consequent no stronger (and perhaps weaker). And the Cooperation Argument can easily be modified around this weaker principle, for one merely needs to suppose that in an SCS: if I rationally ought to adopt the reciprocation intention, then I would do so; if circumstance *c* were to come about, then I would be rationally obliged to believe it had, and I would believe so; and if nothing relevant to the adoption had indeed changed save what must have changed with the coming about of circumstance c, then I would still have the reciprocation intention.

The more plausible claim deserving to be of primary focus in this dispute is, then, principle (G*). Yet this principle merges two central ideas which are, I believe, best kept separate. On the one hand, there is a general view about the relation between rational intention adoption, rational non-reconsideration, and rational action:

(*) If I rationally ought to adopt an intention to do *x* in circumstances *c*, and I do so, and if *c* come about, I rationally ought to believe that *c* has come about, and I do, and *if I rationally ought not to reconsider the intention*, and I still have this intention, then I am rationally permitted to carry out *x*.

On the other hand, there is a specific view about the conditions under which rational reconsideration is appropriate:

(G) If nothing relevant to the adoption of the intention to do *x* in circumstances *c* is changed save what must be changed with the coming about of *c*, then *I rationally ought not to reconsider my intention.*

General principle (*) and the specific claim (G) obviously entail the weakened form, (G*), of Gauthier's Claim. This means that if one wants to circumvent the Deterrence Objection, then one will need to reject principle (G*), and so one will need to reject either (*) or (G). But which one?

A THEORY OF RATIONAL RECONSIDERATION

My central claim in this paper is that we may reformulate the Cooperation Argument by replacing Gauthier's Claim with principle (*), and supplementing this principle with some theory of rational reconsideration other than the defective (G).[17] There are any number of such theories, but in the remainder of the paper I want to examine the implications of supplementing principle (*) with what I shall call a deontological theory of rational reconsideration.[18]

The idea behind the **deontological theory** of rational reconsideration is that when significant stakes are involved, one should reconsider an intention, or not, depending on whether there is any relevant new information regarding that intention. More formally, we may take the following to be a general instance of this idea:

(D) (a) If significant stakes are involved and relevant new information is available about the outcome of the intention to do *x* in circumstances *c*, then I rationally ought to reconsider the intention. (b) If significant stakes are involved and no relevant new information is available about the outcome of the intention, then I rationally ought not to reconsider the intention.

I call this a deontological theory since the rationality of reconsideration is determined not by the value of the outcome of any act or disposition of reconsideration, but rather by a factor—the presence or absence of relevant new information—independent of such value.

Ironically, one motivation for a view such as this comes from Kavka himself.[19] To see how, we begin by noting that Gauthier claims the rational agent is the one who takes the big picture in his aim to fulfill his values, and is driven by plans and intentions. Says Gauthier: "[t]he fully rational actor is not the one who assesses her actions from now but, rather, the one who subjects the largest, rather than the smallest, segments of her activity to primary rational scrutiny, proceeding from policies to performances, letting assessment of the latter be ruled by assessment of the former."[20] Kavka disagrees, and thinks that policies and performances require separate evaluation, but he does admit that "there may be something to"[21] this wider segments view, and that there are clear advantages of agents acting according to rules, plans, and policies, than on a case-by-case basis. Nevertheless, Kavka believes that

> our normal view of rationality also implies being prepared to change previously formulated plans or intentions when there are significant stakes involved and relevant new information about the outcome is available. This is precisely the situation that arises when deterrence fails in an SDS. There is much harm to be done by retaliation, and the benefit that motivated formation of the intention to retaliate—prevention of the offence—is now unobtainable.[22]

Kavka suggests, in this passage, that when there are significant stakes involved, and even if it is initially rational to adopt an intention to *x* when *c*, and condition *c* has come about, then it may be irrational not to reconsider, if relevant new information about the outcome is available—in this case, that the benefit which motivated the formation of the intention is not now available. Kavka suggests that our normal view of rationality seems to include (D)(a).

Unfortunately, Kavka says nothing explicit about what our normal view of rationality implies when there are significant stakes involved but there is not relevant new information. Our 'normal view' of rationality provides a sufficient condition for when one ought to reconsider, and seems to be that if there are significant stakes involved (call this condition 's') and there is new relevant information ('r'), then one ought to reconsider. But what does our 'normal view' say about the conditions under which one ought *not* to reconsider? There are three possibilities within the spirit of this view: (1) if there are significant stakes involved and no relevant new information ('s and not-r'), then one ought not to reconsider; (2) if there are no significant stakes involved and relevant new information ('not-s and r'), then one ought not reconsider; and (3) if either there are no significant stakes involved or no relevant new information ('not-(s and r)'), then one ought not reconsider. (2) is obviously implausible, and (3) stronger than it needs to be for our present discussion. (1) is the weakest principle relevant to the present discussion. We can speculate, then, that our normal view of rationality commits us to something like (1)—that is, (D)(b).

Kavka also does not say very much about how we are to understand the notion of 'relevant new information' about the outcome of an intention.

To motivate an idea of this notion which will be adequate for the purposes of this paper, consider the following example. Suppose I want to meet you to discuss some important matters, and I am to decide whether to go to Hyde Park to do so. Amongst other things, I expect, or have strong reason to expect, that were I to decide to go to Hyde Park, then (I would call you and) you would meet me there within the hour. I decide to go to Hyde Park (I call you up, buy the appropriate ticket for the train there, and so on). Two things might now happen. In the first case, while waiting on the platform at the station, you might give me a call on my mobile phone, and confirm you are now at Hyde Park. Clearly, the fact that you are now at Hyde Park does not constitute relevant new information about the outcome of my decision. It may be relevant information, but it is not new, since at the time of making the decision, I expected, or had strong reason to expect, that this is precisely what would happen. According to our normal view of rationality, this fact is no reason to reconsider my decision to go to Hyde Park. In the second case, you might give me a call and indicate you have been delayed, and cannot be at Hyde Park for some time. Clearly, the fact that you have been delayed does constitute relevant new information about the outcome of my decision. It is obviously relevant information, and it is new, since at the time of making the decision, I expected, or had strong reason to expect, that this is precisely what would not happen. According to our normal view of rationality, this fact is reason to reconsider my decision. (It is not, necessarily, a reason not to go to Hyde Park. The delay will be long, but the matters to be discussed are important. I need to think about it.)

We may generalize from this example. Suppose I expect, or have strong reason to expect, that were I to adopt a certain intention, then p would be the case, and suppose I do adopt this intention. Then (i) if p becomes (or remains) the case, and I believe so, then the fact that p is *not* relevant new information about the outcome of the intention, and (ii) if not-p becomes (or remains) the case, and I believe so, then the fact that not-p *is* relevant new information about the outcome of the intention. In short, if things turn out as I expected they would when I adopted the intention, then I have no relevant new information; and if they do not, then I do.

This account is consistent with what Kavka has to say on the matter. He suggests that the relevant new information occurring when deterrence fails in an SDS is that "the benefit that motivated formation of the intention to retaliate—prevention of the offence—is now unobtainable."[23] The account in the previous paragraph implies that this is indeed so. For in an SDS, I expect, or have strong reason to expect, that were I to adopt the deterrent intention, then prevention of the offence would be obtainable (and, indeed, obtained), and, in an SDS, I have adopted this intention. And so, when deterrence fails in an SDS, the fact that prevention of the offence is no longer obtainable is, as Kavka suggests, relevant new information.

III. Rational Cooperation, Irrational Retaliation

The key step in reformulating the Cooperation Argument is to replace Gauthier's Claim by principle (*), and to supplement this principle, in particular, by theory (D) of rational reconsideration. In this paper, I will not attempt to defend either of these claims in detail, apart from noting that both seem plausible on first encounter—principle (*) asserts what appears to be a platitudinous relation between the rationality of intention, reconsideration, and action, and theory (D) is a view of rational reconsideration which even Kavka seems to find attractive. Of course, in itself this will not move those, such as Kavka, who think SDSs provide counterexamples to so-called 'bridging principles' such as (*).[24] What should move them, though, is the fact that even if SDSs provide counterexamples to other bridging principles, they provide no counterexample to principle (*). For in this final section I will argue that, under principle (*) and theory (D), cooperation is rational in SCSs, but retaliation irrational in SDSs.

RATIONAL COOPERATION, IRRATIONAL RETALIATION

SDSs and SCSs are indeed very similar in many respects. To get you to acquiesce in an SDS, it is necessary and very likely sufficient for me to adopt the deterrent intention; to get you to cooperate in an SCS, it is necessary and very likely sufficient for me to adopt the reciprocation intention. Thus it is rational for me in an SDS to adopt the deterrent intention; it is rational for me in an SCS to adopt the reciprocation intention. Retaliating in an SDS, however, would not be expected-value maximizing, nor would cooperating in a SCS.

SDSs and SCSs, though, are crucially different in one respect. To see this, concentrate on what theory (D) has to say about whether the relevant intention should be reconsidered.

On the one hand, in an SDS where deterrence has failed, theory (D) implies that I ought to reconsider my deterrent intention, and so, presumably, that I ought not retaliate.[25] In an SDS, the relevant intention is an intention to retaliate (= x) in the circumstances that you fail to acquiesce first (= c). This means that in an SDS, if the relevant circumstance (namely, that you fail to acquiesce) comes about, then I will know that things have *not* turned out as I expected (since, recall, my adopting the deterrent intention is supposed to make it *very likely* that you *will* acquiesce), and so the stakes are high and relevant new information about the outcome of adopting the intention is available. This means we may infer from (D) that I rationally ought to reconsider the deterrent intention in an SDS. We can agree with Kavka that our normal view of rationality implies that when things are going exactly *contrary* to how we thought they would when we devised our plans, then it is indeed irrational not to reconsider those plans, and it may be irrational to act on them.

On the other hand, in an SCS where my promise has succeeded, theory (D) implies that I ought *not* to reconsider my reciprocation intention, and so, in conjunction with (*), that I am rationally permitted to cooperate. In an SCS, the relevant intention is an intention to cooperate (= x) in the circumstances that you cooperate first (= c). This means that in an SCS, if the relevant circumstance (namely, that you cooperate) comes about, then I will know that things *have* turned out as I expected (since, recall, my adopting the cooperation intention is supposed to make it *very likely* you *will* cooperate), and so the stakes are high and no relevant new information about the outcome of adopting the intention is available. This means we may infer from (D) that it is rational not to reconsider the reciprocation intention in an SCS. We can insist that our normal view of rationality also implies that when things are going exactly as we thought they would when we devised our plans, then it is indeed rational not to reconsider those plans, and therefore rational to act on them.

But is there not other information—relevant and new—in an SCS after you have cooperated?[26] Consider just two possibilities. Before you cooperated, your cooperation was dependent on the presence of my intention; after, it is not. And doesn't the fact that your cooperation no longer depends on my intention constitute relevant new information about the outcome of the intention? Before you cooperated, my fields were not harvested; after, they were. And doesn't the fact that my fields are now harvested constitute relevant new information about the outcome of the intention? If this is so, then theory (D) implies I rationally ought to reconsider, and so, presumably, that I ought not to cooperate.

In response, I claim that while these two pieces of information are certainly relevant to the case at hand, they do not constitute 'new' information in the appropriate sense. Of course, I agree that after you have cooperated, your cooperation becomes independent of my intention, and my fields become harvested. But this is precisely what I expected, or had strong reason to expect, before I adopted the intention. This is so because (as is agreed by all) I expected, or had strong reason to expect, that were

I to adopt the intention, then you would indeed cooperate, and (on independent grounds) I believed, or had strong reason to believe, that after you cooperated, your cooperation would become independent of my intention, and my fields would be harvested. Since I expected, or had strong reason to expect, these things to come about, the fact that they do come about is hardly new information.

One must not confuse two different senses of 'new' information. Consider again the Hyde Park example. In the first case, you give me a call on my mobile phone, and confirm you are now at Hyde Park. This is, in one sense, new information: where formerly you were not at Hyde Park, now you are. Something has changed (though my expectations have been fulfilled), but I have no reason to reconsider my intention to go to Hyde Park. So too I have no reason to reconsider my intention to cooperate. In the second case, you call to tell me that you will be significantly delayed. This is, in another sense, new information: formerly you were not at Hyde Park, and even now you are not at Hyde Park, but this is contrary to what I expected. My expectations have not been fulfilled (though your location has not changed), and I have reason to reconsider my intention. So too I have reason to reconsider my intention to retaliate. According to theory (D), and according to intuition, reconsideration should be prompted not by the fact that something has changed, but rather by the fact that expectations have not been fulfilled.

CAN THE DETERRENCE OBJECTION BE REFORMULATED?

SDSs and SCSs are not, then, relevantly similar. But can SDSs be changed so that (i) they are relevantly similar to SCSs, and yet (ii) they still provide a strong objection to the (reformulated) Cooperation Argument?

In general, SDSs will be relevantly similar to SCSs only if the value of the status quo is close to the value of the outcome of retaliation, or the probability, Δ, is close to zero that I would retaliate were I to adopt the intention and you fail to acquiesce (see Figure 2). To see this, note, first, that SDSs are relevantly similar to SCSs only if (1a) it maximizes expected-value to adopt the deterrent intention, and (1b) the probability, μ_r, is close to one that you would not acquiesce were I to adopt the deterrent intention. The justification for (1a) is that it maximizes expected-value in an SCS to adopt the reciprocation intention, and for (1b) is that this would mean that your failure to acquiesce would not constitute relevant new information. Second, it maximizes expected-value to adopt the deterrent intention in an SDS if and only if $\mu_r < (t{-}d)/[(t{-}d) + \Delta(d{-}h)]$.[27] Hence, third, SDSs are relevantly similar to SCSs only if (1a') $\mu_r < (t{-}d)/[(t{-}d) + \Delta(d{-}h)]$, and (1b') $\mu_r \approx 1$. And from these two facts it follows, as required, that either the value of the status quo is close to the value of the outcome of retaliation ($d \approx h$), or the probability is close to zero that I would

retaliate were I to adopt the intention and you fail to acquiesce ($\Delta \approx 0$). It turns out, however, that the Deterrence Objection is strong in neither of these cases.

On the one hand, if the value of the status quo is close to the value of the outcome of retaliation, then the Deterrence Objection is question-begging. For in this case retaliating and doing nothing are actions with roughly the same value. This means that retaliating must consist of some trivial non-expected-value maximizing action, which it would be question-begging to insist is irrational. In this first case, then, the Deterrence Objection would not be a strong one.

On the other hand, if the probability is close to zero that I would retaliate were I to adopt the intention and you fail to acquiesce, then it may seem the Deterrence Objections stands. Retaliation in such deterrence situations remains a horrendously destructive act, and such deterrence situations are relevantly similar to SCSs. In this type of case, then, it is rational to adopt the deterrent intention (this is in part the definition of an SDS), rational to believe it has failed (these are the deterrence situations of current interest), rational, according to (D), not to reconsider (since your failure to acquiesce was expected), and so, according to (*), rationally permitted to retaliate. But it is clearly not rational to retaliate. Thus, if we add a fifth condition to the definition of an SDS—namely, that the probability is close to zero that I would retaliate after a failed attempt at deterrence—then it seems the Deterrence Objection can itself be reformulated against principles (D) and (*), and thus remains an objection to the reformulated Cooperation Argument.

The first point to be made against this condition is that, even if it is acceptable to add, it does not threaten the reformulated Cooperation Argument. The relevant part of principle (*) states that "If ..., and ..., and if I rationally ought not to reconsider the intention, and *I still have this intention*, then I am rationally permitted to carry out *x*". Granted, theory (D) implies that I rationally ought not reconsider the intention; but in these modified deterrence situations, I would very likely lose the intention anyway. For if the probability is close to zero that I would retaliate were my deterrence to fail, then the probability must be close to one that I would lose the intention, since if I still had the deterrent intention, and came to believe you had failed to acquiesce, then the probability I will retaliate would certainly not be close to zero. This means it does *not* follow, according to (*), that I am rationally permitted to retaliate.[28]

The second point to be made against this fifth condition is that it is in any case unacceptable to add. The first two conditions of the definition of an SDS imply that, to get you to acquiesce, it is necessary and very likely sufficient for me to adopt the deterrent intention. Faking the intention is not enough, and this must be because you can tell whether or not I have the intention. The putative fifth condition, however, states that there is little chance I would actually carry out my intention, were you not to acquiesce. But since you can tell whether or not I have the intention, you can also probably tell whether or not I am disposed to act on it. (It seems unmotivated to suppose you can tell what my intentions are, without also being able to tell what my dispositions are.) Thus, the implications of the first two conditions, and the putative fifth, are (i) that if I have the deterrent intention then you will probably acquiesce, but (ii) that you know I would not carry out the intention were you not to acquiesce. In a word, the implication of these three conditions is that you are exceedingly dim-witted. Since such deterrence situations are uninteresting, and discussing them is probably not what Kavka had in mind, it is unacceptable to suppose that I would not carry out my deterrent intention. In this second case, then, the Deterrence Objection would also not be a strong one.

If we do not suppose I would not carry out my intention, then the idea of an SDS remains interesting, though not, of course, because it provides the basis for an objection to the rationality of cooperation in SCSs. Deterrence Situations cannot be changed so that they are relevantly similar to SCSs, and yet remain strong objections to the (reformulated) Cooperation Argument.

IV. Conclusion

The Cooperation Argument can be reformulated, then, by replacing its second premise (Gauthier's Claim) by the conjunction of (*) and (D). And such a reformulation would be attractive indeed. First, we get to keep the principle (*)—a principle asserting what seems to be a platitudinous relation between rational intention adoption, rational non-reconsideration, and rational action. Second, we get to keep theory (D)—a theory of rational reconsideration which even Gregory Kavka admits "there may be something to." Third, with the argument reinstated, and other objections not withstanding, we may avail ourselves of its conclusion that it might be rational to cooperate even though it does not maximize expected-value to do so—an important conclusion if Gauthier's attempt to secure a rational morality is to succeed. And, finally, even with the argument reinstated, we are not committed to the conclusion that it is rational to retaliate—a conclusion many find very unattractive, and one which I believe explains the force of the Deterrence Objection. The reformulation provides all we could ever want, and more.[29]

University of Wollongong
Wollongong, New South Wales

NOTES

[1] D. Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986), pp. 157–189. Solely for the sake of convenience, in this paper I offer the following definition of the expected-value, EV(A), of an action A: $\mathrm{EV(A)} = \Sigma_i\, \mathrm{P}(\mathrm{A}\Box\!\!\rightarrow\mathrm{O}_i).\mathrm{V}(\mathrm{A}\&\mathrm{O}_i)$. This is equivalent to the definition of U-expectation in A. Gibbard and W. Harper, "Counterfactuals and Two Kinds of Expected Utility," in C. A. Hooker, J. J. Leach, and E. F. McClennen, *Foundations and Applications of Decision Theory: Volume I* (Dordrecht: Reidel, 1978), pp. 125–62. Any other formal definition of 'expected-value' would do as well for the purposes of this paper. In addition, I sometimes use the term 'maximizing action' as short for 'expected-value maximizing action'

[2] J. Harsanyi, "Review of 'Morals by Agreement'," *Economics and Philosophy* 3 (1987), pp. 339–373, esp. section 3; D. MacIntosh, "Two Gauthiers?" *Dialogue* 28 (1989), pp. 43–61, esp. sections 3, 4, and 5.

[3] The possibility of this type of objection has been suggested to me by a number of people, including John Broome and Galen Strawson. I take responsibility, however, for its formulation. See also D. MacIntosh, "Retaliation Rationalized: Gauthier's Solution to the Deterrence Dilemma," *Pacific Philosophical Quarterly* 72 (1991), pp. 9–32.

[4] On the assumption of transparency, and that of translucency, see, for example, A. Nelson, "Economic Rationality and Morality," *Philosophy and Public Affairs* 17 (1988), p. 160, R. J. Arneson, "Locke versus Hobbes in Gauthier's Ethics," *Inquiry* 30 (1987), p. 309, D. Copp, "Contractarianism and Moral Skepticism," and G. Sayre-McCord, "Deception and Reasons to be Moral," in P. Vallentyne, *Contractarianism and Rational Choice: Essays on Gauthier's 'Morals by Agreement'* (New York: Cambridge University Press, 1991), pp. 220–221, and 191–195 respectively.

[5] G. Kavka, "Review of 'Morals by Agreement'," *Mind* 96 (1987), p. 120. Others making this type of objection include G. Harman, "Rationality in Agreement: A Commentary on Gauthier's 'Morals by Agreement'," *Social Philosophy and Policy* 5 (1988), pp. 1–16; D. Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1986); D. Lewis, "Devil's bargains and the Real World", in D. Maclean (ed.), *The Security Gamble: Deterrence Dilemmas in a Nuclear Age* (Totowa: Rowman & Allanheld, 1984), pp. 141–154; and S. Darwell, "Rational Agent, Rational Action," *Philosophical Topics* 14 (1986), pp. 33–57.

[6] This example is a slightly modified version of one occurring in D. Gauthier, "Why Contractarianism?" in P. Vallentyne, *Contractarianism and Rational Choice: Essays on Gauthier's 'Morals by Agreement'*, p. 24. See also D. Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), p. 7.

[7] Where: I=adopt intention; C=cooperate. I assume that if I were not to adopt the intention then you would not cooperate, and help me with my harvest, and I would in turn not cooperate, and help you with yours. Further, I assume that if I do adopt the intention, but you do not cooperate, then I will not cooperate in return. Either way, poor harvests for each would be the result.

[8] In this paper I say that the agents in question 'adopt' the relevant intentions, and leave as an open question the nature of this adoption, be it through an act of will, more indirectly, or perhaps by some other means. Whether agents can adopt intentions to perform non-maximizing actions is, as I have already indicated, not the issue for this paper—I shall assume they can.

[9] A simple calculation shows this. $\mathrm{EV(I)} = \mathrm{P}(\mathrm{I}\Box\!\!\rightarrow\mathrm{C}).\mathrm{EV}(\mathrm{I}\&\mathrm{C}) + \mathrm{P}(\mathrm{I}\Box\!\!\rightarrow -\mathrm{C}).\mathrm{EV}(\mathrm{I}\&-\mathrm{C}) = (1-\mu_\mathrm{C}).\mathrm{EV}(\mathrm{I}\&\mathrm{C}) + \mu_\mathrm{C}.d \geq (1-\mu_\mathrm{C}).c + \mu_\mathrm{C}.d$, since $\mathrm{EV}(\mathrm{I}\&\mathrm{C}) \geq c$ and letting $\mu_\mathrm{C} = \mathrm{P}(\mathrm{I}\Box\!\!\rightarrow -\mathrm{C})$, and assuming $\mathrm{P}(\mathrm{I}\Box\!\!\rightarrow \mathrm{C}) = 1-\mathrm{P}(\mathrm{I}\Box\!\!\rightarrow -\mathrm{C}) = 1-\mu_\mathrm{C}$. Continuing, $(1-\mu_\mathrm{C}).c + \mu_\mathrm{C}.d > (1-\mu_\mathrm{C}).d + \mu_\mathrm{C}.d = d = \mathrm{EV}(-\mathrm{I})$, if $\mu_\mathrm{C}<1$. Hence, if $\mu_\mathrm{C}<1$, then $\mathrm{EV(I)} > \mathrm{EV}(-\mathrm{I})$.

[10] I have adopted the following form of the definition solely for the sake of my opponent's argument, to make Special Cooperation Situations as similar as possible to Special Deterrence Situations.

[11] D. Gauthier, "Afterthoughts", in D. MacLean (ed.), *The Security Gamble: Deterrence Dilemmas in a Nuclear Age* (Totowa: Rowman & Allanheld, 1984), p. 159. It turns out Gauthier no longer endorses this principle (see "Assure and Threaten," *Ethics* (forthcoming)); however it is still useful as a starting point for my own discussion.

[12] This example is a modified version of ones occurring in a number of places. See, for example, T. C. Schelling, *The Strategy of Conflict* (London: Oxford University Press, 1963), and D. Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), pp. 20 ff.

[13] Typically, of course, we say that I 'deter' you only if I make a threat to prevent you from producing some harm (for me), and not—as in the case at hand—if I make a threat to induce you to produce some benefit (for me). In this paper, I will say that I 'deter' you only if I make a threat to get you to perform some action with greater expected-value (for me) than the action you would have otherwise performed. This covers both preventing you from producing some harm, and inducing you to produce some benefit.

[14] Where I=adopt intention; A=acquiesce; R=retaliate. I assume that if I were not to adopt the intention then you would not acquiesce, and give me your oil fields, and I would in turn not retaliate, and start a war with you. Further I assume that if I do adopt the intention, but you do acquiesce, then I will not retaliate in return. In the first case the status quo would result, and in the second I would end up with your oil fields.

[15] A simple calculation shows this. $EV(I) = P(I\Box\!\!\rightarrow A).EV(I\&A) + P(I\Box\!\!\rightarrow -A).EV(I\&-A) = (1-\mu_r).t + \mu_r.EV(I\&-A) \geq (1-\mu_r).t + \mu_r.h$, since $EV(I\&-A) \geq h$ and letting $\mu_r = P(I\Box\!\!\rightarrow -A)$, and assuming $P(I\Box\!\!\rightarrow A) = 1-P(I\Box\!\!\rightarrow -A) = 1-\mu_r$. Continuing, $(1-\mu_r).t + \mu_r.h = t-\mu_r.(t-h) > t-(t-d) = d = EV(-I)$, if $\mu_r < (t-d)/(t-h)$. Hence, if $\mu_r<(t-d)/(t-h)$, then $EV(I) > EV(-I)$.

[16] Special Deterrence Situations are closely related to situations of the same name Kavka describes in "Some Paradoxes of Deterrence," *Journal of Philosophy* 75 (1978), pp. 285–302. This paper is reprinted, with alterations, in G. Kavka, *Moral Paradoxes of Nuclear Deterrence* (New York: Cambridge University Press, 1987), pp. 15–32. In particular, the definition of an SDS in the later version is slightly different, and I will base my own definition of an SDS only on the later version of the paper.

The definition of an SDS in the text differs in a number of places from the one Kavka offers. (1) Kavka is concerned with the *morality* of deterrence, while I am concerned with the *rationality* of deterrence. I believe—and Kavka agrees ("A Paradox of Deterrence Revisited", in *Moral Paradoxes*, pp. 43 ff.)—that this makes no difference to the validity of the relevant arguments. (2) Kavka says it is likely the agent must intend (conditionally) to apply a harmful sanction to innocent people, if an extremely harmful and unjust offence is to be prevented; while I say the agent must intend (conditionally) to apply a sanction harmful to themselves if an offence harmful to themselves (namely, the other not cooperating) is to be prevented. Kavka requires it to be 'likely' to be necessary, while I say it 'must' be necessary. This simplifies the discussion, and (if anything) strengthens Kavka's case. (3) Kavka says the agent would have conclusive moral reasons not to apply the sanction if the offence were to occur, while I say that the expected-value of not retaliating is greater than that of retaliating, even if the other were not to acquiesce. He would suppose, in the rationality case, that the agent in question would have conclusive (rational) reasons against retaliating, or, in short, that they rationally ought not to retaliate. But to provide such a condition as part of a characterization of SDSs is to beg the question against those—such as Gauthier—who are concerned to argue that it could be rational to perform a non-maximizing action. A non-question-begging final condition would be that the agent in question values the outcome of not retaliating to that of retaliating, and this is the type of condition I have adopted.

[17] See M. Bratman, *Intentions, Plans and Practical Reason* (Cambridge, Harvard University Press, 1987), section 6.6, pp. 101–106, for a discussion of the view that it might be rational to reconsider even though nothing has changed save what must have changed with the coming about of the relevant circumstance. I have no space in this paper to discuss Bratman's own thoughts concerning reconsideration.

[18] I discuss two other theories elsewhere. The rule-consequentalist theory, (R), states that one ought to reconsider (or not) if and only if expected-value maximizing habits of reconsideration would have one (not) reconsider. Under theory (R), one rationally ought not reconsider a failed deterrent intention in SDSs where, in order to get you to acquiesce, it is necessary and very likely sufficient for me to adopt the deterrent intention and be disposed to carry it out. The act-consequentalist theory, (A), states that one ought to reconsider (or not) if and only if expected-value maximizing act is (not) to reconsider. Under theory (A), one rationally ought not reconsider a failed deterrent intention in SDSs where I am the sort of person who judges (rightly or wrongly) that the fact you have failed to acquiesce is, in and of itself, a conclusive reason for me to retaliate. For further details, see my "Retaliation Rationalised?", unpublished m.s.

[19] G. Kavka, "The Paradox of Deterrence Revisited," in his *Moral Paradoxes of Nuclear Deterrence* (New York: Cambridge University Press, 1987), pp. 45–6.

[20] D. Gauthier, "Deterrence, Maximization, and Rationality," *Ethics* 94 (1984), p. 488. See also D. Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986), pp. 157–189.

[21] Kavka, 'The Paradox of Deterrence Revisited,' p. 45

[22] Kavka, 'The Paradox of Deterrence Revisited,' pp. 45-46.

[23] Kavka, 'The Paradox of Deterrence Revisited,' p. 46.

[24] Kavka discusses three such 'bridging principles' in *Moral Paradoxes of Nuclear Deterrence* (New York: Cambridge University Press, 1987), pp. 15–32. These are: the *Wrongful Intentions Principle*—to form an intention to do what one knows to be wrong is itself wrong; the *Right-Good Principle*—doing something is right if and only if a morally good person would do the same thing in the given situation; and the *Virtue Preservation Principle*—that it is wrong to deliberately lose (or reduce the degree of) one's moral virtue. Principle (*) is one more bridging principle—though, as I argue, not one susceptible to the paradox of deterrence.

[25] In this paper, I assume (*), and so assume (abbreviating somewhat) that if I ought not to reconsider the relevant intention, then I am permitted to act on it. I also assume—solely for the sake of my opponent's argument—that if I ought to reconsider the relevant intention, then I ought not act on the intention. This is, of course, not generally true, since even if one ought to reconsider an intention, it may be that one will be re-confirmed in that intention, and thus that one is indeed permitted to act on it.

[26] I would like to thank Michael Smith for bringing this point to my attention. The formulation of the objection, however, is my own.

[27] A simple (if tedious) calculation shows this. (In this note, I write 'μ' for 'μ_r'.) $EV(-I) = d$, straightforwardly. $EV(I) = P(I \Box\!\!\rightarrow A).EV(I\&A) + P(I \Box\!\!\rightarrow -A).EV(I\&-A) = (1-\mu).t + \mu.EV(I\&-A)$, since $EV(I\&A) = t$, and letting $\mu = P(I\Box\!\!\rightarrow -A)$, and assuming $P(I \Box\!\!\rightarrow A) = 1-P(I \Box\!\!\rightarrow -A) = 1-\mu$. Now $EV(I\&-A) = P(I\&\text{-}A\Box\!\!\rightarrow R).EV(I\&\text{-}A\&R) + P(I\&\text{-}A \Box\!\!\rightarrow -R)$. $EV(I\&\text{-}A\&-R) = \Delta.h + (1-\Delta).d$, since $EV(I\&\text{-}A\&R) = h$, $EV(I\&\text{-}A\&-R) = d$, and letting $\Delta = P(I\&\text{-}A \Box\!\!\rightarrow R)$, and assuming $P(I\&\text{-}A \Box\!\!\rightarrow -R) = 1-P(I\&\text{-}A \Box\!\!\rightarrow R) = 1-\Delta$. Substituting, $EV(I) = (1-\mu)t + \mu\Delta h + \mu(1-\Delta)d$. It follows that $EV(I) > EV(-I)$ iff $(1-\mu)t + \mu\Delta\mathrm{h} + \mu(1-\Delta)d > d$, iff $\mu < (t-d)/[(t-d) + \Delta(d-h)]$.

[28] Of course, this does leave the cases in which I rationally ought not to reconsider, and I do indeed not reconsider, even though the chance I would reconsider is high. Principle (*) applies to such cases, and so it follows from this principle that I am rationally permitted to retaliate. There is no choice in this case for the proponent of the reformulated argument but to bite the bullet, and accept this conclusion, though this should not be too difficult, since these cases are, by definition, ones that are very much more unlikely than SDSs themselves. See also the second point I make against the fifth condition.

[29] I would like to thank John Broome, Robert Dunn, Nicole Gerrand, Frank Jackson, Julian Lamont, Peter Menzies, Philip Pettit, Michael Smith, Galen Strawson, and Susan Wolf for their helpful comments.

PACIFIC PHILOSOPHICAL QUARTERLY

Contents of
VOLUME 74
(1993)

PUBLISHED QUARTERLY BY
THE SCHOOL OF PHILOSOPHY

UNIVERSITY OF SOUTHERN CALIFORNIA

PACIFIC PHILOSOPHICAL QUARTERLY
Journals Marketing Manager
Blackwell Publishers
108 Cowley Road
OX4 1JF UK
or 238 Main Street, Cambridge, MA 02142, USA

Subscription Prices 1994	North America	UK/Europe	Rest of World
Institutions	$59.00	£45.50	£50.50
Individuals	$33.50	£21.00	£24.00