Deterrence, Maximization, and Rationality*

David Gauthier

I

Is deterrence a fully rational policy? In our world deterrence works—sometimes. But in a more perfect world, in which actors rationally related their choices to their beliefs and preferences, and in which those beliefs and preferences were matters of common knowledge, could deterrence work? Some say no.[1] Others hold a conception of rationality that would commit them to saying no, were they to consider the issue.[2] I say yes. Deterrence can be part of a fully rational policy. I propose to demonstrate this.

At the heart of a deterrent policy is the expression of a conditional intention. An actor A expresses the intention to perform an action x should another actor B perform an action y. If B would do y did A not express her intention, then we may say that A's expression of intention deters B from doing y. In expressing her intention as part of a deterrent policy, A seeks to decrease the probability of B's doing y by increasing his estimate of her conditional probability of doing x should he do y.

We need better labels than x and y if our talk about deterrence is to be perspicuous. In at least some situations, A's deterrent intention is *retaliatory;* A expresses the intention to retaliate should B do y. So let us call x *retal*. And what A seeks to deter is an action that would advantage B in relation to A; let us then call y *advant*. We shall then say that an actor A expresses the intention to *retal* should another actor B *advant*.

A seeks to affect B's estimate of her conditional probability of *retal* should he *advant*. Why does she expect her expression of conditional intention to have this effect? Let us suppose that A and B are rational; on the received view of rationality, an actor seeks to maximize expected

* This paper was prepared for delivery at a conference on "Nuclear Deterrence: Moral and Political Issues," sponsored by the Center for Philosophy and Public Policy, University of Maryland at College Park. It will appear in *The Security Gamble: Deterrence Dilemmas in the Nuclear Age*, edited by Douglas MacLean, Maryland Studies in Public Philosophy (Totowa, N.J.: Rowman & Allanheld, in press).

1. One who says no is Jonathan Schell, *The Fate of the Earth* (New York: Alfred A. Knopf, Inc., 1982), pp. 201–4.

2. Among these others are game theorists who insist that strategic rationality demands perfect equilibria.

Ethics 94 (April 1984): 474–495
 0014-1704/84/9403-0007$01.00

utility, the fulfillment of her preferences given her beliefs. If A expects to affect B's estimate of what she will do, then she must expect to affect his beliefs about her preferences and/or beliefs. Or so it seems.

A wants to deter B from *advant*. She believes that B is less likely to *advant* if he expects her response to be *retal* than if he expects a different response, *nonretal*. She therefore expresses the intention to *retal* should he *advant*. For this to affect B, it would seem that he must take her expression of intention to indicate her preference for *retal* over *nonretal*, given *advant*. Perhaps A does have this preference and so seeks to inform B that she prefers *retal*. Perhaps A does not have this preference but seeks to deceive B into supposing that she prefers *retal*. But in either case the deterrent effect of her expression of intention would seem to require that B be initially uninformed, or at least uncertain, about her preference. Were he informed of her preference, then his estimate of her conditional probability of choosing *retal* should he *advant* would be unaffected by any claim she might make about her intention.

But is this so? Must the actor to be deterred be initially uncertain about the preferences of the would-be deterrer? Let us consider the matter more closely. We suppose that B knows A's preferences between *retal* and *nonretal*, given *advant*. If she prefers *retal*, then his knowledge should suffice to deter him from *advant*, supposing that his preferences are such that he can be deterred at all. A needs no deterrent policy. If she prefers *nonretal*, then how can her expression of the conditional intention to *retal* should he *advant* be credible? How can it affect his estimate of what she will do?

First we might suppose that, although A prefers *nonretal* to *retal* ceteris paribus, yet she also prefers being a woman of her word. She may value sincerity directly, or she may find it instrumentally useful to her. In expressing her intention to *retal* should B *advant*, she stakes her reputation for being a woman of her word, and B, knowing or believing this, realizes that by expressing her intention she has transformed the situation. She prefers *nonretal* to *retal*, but she also prefers honoring a commitment leading to *retal* to dishonoring a commitment even if it brings about *nonretal*. Her expression of conditional intention does not affect her preferences but brings a different set into play and so affects B's estimate of the utilities of the courses of action open to her should he *advant*.

Second, A may be imperfectly rational, unable fully to control her behavior in terms of her considered preferences. If B *advant*s, then her cool preference for *nonretal* may be overcome by anger, or rage, or panic, so that she may *retal*. In this case we should no doubt say, not that A expresses a conditional intention to *retal*, but rather that she expresses a warning that she will, or may, find herself choosing *retal* should he *advant*. Fortunately for A, her inability to control her behavior stands her in good stead, enabling her to deter, or at least to seek to deter, B from *advant* by warning him of her probable folly should he do it. Such

an inability may seem suspect, as altogether too convenient, making us hesitant to accept this apparent mode of deterrence at face value.

Third, A's expression of intention may not stand alone but may activate forces themselves beyond her control, which may make *nonretal* less desirable, or *retal* more desirable, than would otherwise have been the case. Perhaps A has made a side bet which she loses should she fail to abide by her stated intention, or perhaps she has insured herself against the costs of having to carry out what otherwise would be an unprofitable course of action. And fourth, in expressing her intention, A may also delegate her power to choose; some other person, or some preprogrammed device, capable of ignoring her preferences, will ensure that if B *advants*, *retal* will ensue. These complicating cases will play no part in our discussion. My interest in this paper is in deterrent policies that do not call into play external factors no longer within the actor's control.

My interest is also in genuine expressions of intention, and not in warnings. No doubt we are not always in such control of our actions that our cool, long-term, considered preferences prevail. But as I have noted, there is something suspect about arranging to gain from this lack of control, about extracting rational advantage from seeming irrationality. I shall consider would-be deterrers who are able to carry out what they intend and who form their intentions on a rational, utility-maximizing basis. A then does not warn B but coolly informs him that she will deliberately *retal* should he *advant*.

And lastly, my interest is not in the provision of deterrent information about preferences. Rather we shall examine situations in which there is no doubt, in the minds of those concerned, that, at least if other things are equal, the would-be deterrer A disprefers *retal* to *nonretal*, should B *advant*.

It would therefore seem that we are left with but one possibility for a deterrent policy among rational persons informed of each other's preferences and beliefs. We must suppose that the would-be deterrer prefers to be a person of her word. A, in expressing her conditional intention, must transform the situation, preferring to abide by her commitment even though, ceteris paribus, she would prefer the outcome of ignoring the commitment. She prefers *nonretal* to *retal*, but having expressed the intention to *retal* given *advant*, she prefers to carry out her intention to ignoring it, should her attempt to deter fail.

Although some deterrent policies may seem to invite this characterization, there are, in my view, insuperable difficulties with it, if we insist firmly on the full rationality of the actors. Of course, since we impose no a priori constraints on the content of preferences, an actor may simply take satisfaction in making commitments which she then carries out. But why would a rational actor choose to make commitments to dispreferred courses of action? Perhaps she finds masochistic satisfaction in making and carrying out such commitments. But if deterrent policies

are rational only for a peculiar variety of masochist, then most real-world examples of such policies survive only because of irrationality. Let us not be so hasty to judge them. I shall suppose that in general, the actor's concern is with the instrumental and not the intrinsic benefits of adhering to an expressed intention. What are these benefits? What does A gain if she actually responds to *advant* by *retal*, having expressed the intention so to respond?

If B *advants*, then A's attempt to deter him has failed. Any gain that would compensate for the cost of *retal* must then derive from further, future consequences of choosing *retal* that extend beyond the particular deterrent situation. Presumably these consequences are the effects of carrying out her expressed intention, on the deterrent value of expressing similar intentions in other situations. If A *retals*, showing that her expression of intention was seriously meant, then future, similar expressions of intention should have a greater effect on others' expectations of what she will do than if she fails to *retal*.

But among fully rational persons is this effect possible? If A is rational, then B rationally expects her to do what she believes will maximize her expected utility. What she has done in the past may provide information about her preferences and beliefs, but we are supposing these to be common knowledge. How then can what A has done affect B's expectation of what she will do in the future? He expects her to maximize her expected utility; how can what she has done affect her expected utility? We are not concerned with behavior that alters the payoffs or outcomes possible for A. If in choosing *retal* A neither informs B about her preferences nor alters the possible outcomes of her future choices, then B has no reason to take what she has done into account in forming his expectations about what she will do in the future. A rational observer, informed of A's preferences, could only interpret her choice of *retal* as a lapse from rationality, in no way affecting expectations about her future choices on the supposition that they will be made rationally.

The only expectation one can rationally form about rational utility-maximizers is that they will seek to maximize expected utility. The only reputation they can rationally gain is the reputation for maximizing expected utility. If carrying out an expressed intention is not itself utility maximizing, then it can have no effect on the expectations of rational and informed persons that would suffice to make it utility maximizing.

To suppose otherwise is to fail to think through the forward-looking implications of maximizing rationality. A utilitarian, dedicated to collective maximization, cannot have reason to keep his promises in order to gain a reputation as a promise keeper among a community of utilitarians, although he may have reason so to act among us nonutilitarians. Similarly, an individual utility maximizer can have no reason to carry out her intentions, in order to gain a reputation as a woman of her word, among a community of informed individual utility maximizers, although she may have reason so to act among less rational persons. We seem then to

have exposed a deep irrationality at the core of deterrent policies. Leaving aside the provision of information about one's preferences, or the issuance of a warning about one's irrationality, or the invocation of factors beyond one's control that would determine one's response, we seem forced to conclude that A cannot expect B to alter his estimate of her conditional probability to *retal* should he *advant*, on the basis of her expressed intention to *retal*, if ceteris paribus she would prefer *nonretal*. And so A cannot expect to decrease the probability of B choosing *advant* by her expression of conditional intention; she is not able to deter, or rationally to attempt to deter, B from *advant*.

II

Or so it would seem. I shall show that things are not what they seem and that it may be rational to adopt a deterrent policy committing one to the performance of a disadvantageous, non-utility-maximizing action should deterrence fail. But before turning to this demonstration, let us pause to entertain the possibility that my argument has been mistaken and that A might have reason to carry out an otherwise disadvantageous expressed intention because of its effect on expectations about her future behavior. It is clear that this can be relevant to the rationality of a deterrent policy only if A is concerned about future deterrence.

Although our analysis of deterrence is intended to apply generally, yet I am particularly concerned with the rationality of deterrent policies in the context of relations among those nations possessing nuclear weapons. More precisely, I am concerned with a policy which has as its core the expressed intention to respond to a nuclear strike with a counterstrike. I shall call this the policy of "nuclear retaliation."

To exemplify this policy and set it in the context of deterrence, let us suppose that one nation—call it the SU—is perceived by another nation—call it the US—to constitute a nuclear threat. The US fears that the SU will launch a nuclear strike, or, perhaps more plausibly, will credibly threaten to launch such a strike should the US refuse some demand or resist some initiative, or, perhaps more plausibly still, will act in some way inimical to the interests of the US that could be effectively countered only by markedly increasing the probability that the SU will launch a nuclear strike. The US seeks to deter the SU from a policy that would or might lead to a nuclear strike, whether unconditionally or as a result of US refusal to acquiesce in or endeavor to counter some SU initiative. To do this, the US announces the intention to resist any SU initiative even if resistance invites a nuclear strike and, should a strike occur, to retaliate even if this provokes full-scale nuclear combat. In talking about the "strike policy" of the SU, and the "retaliatory policy" of the US, I shall intend the policies just sketched. In particular, a strike policy may center on the threat to strike should some demand not be met, and a retaliatory policy may center on the refusal to submit to such a demand even though a nuclear exchange may result.

Now it is possible that the US prefers suffering a nuclear strike to submitting to a demand by the SU. And it is possible that the US prefers retaliating against a nuclear strike, with the prospect then of fighting a nuclear war, to accepting passively a single strike and so, effectively, cutting its nuclear losses by capitulating. But suppose, plausibly, that the consequences of nuclear warfare are such that the US would always prefer less nuclear devastation to more; nevertheless it seeks to deter the SU from a strike policy by expressing the intention to choose its less preferred retaliatory response. It is then engaged in just the type of deterrent policy that we have put rationally in question. And it seems clear that an appeal to future expectations would not here provide ground for altering US preferences in order to defend deterrence in terms of future effects. For the US to claim that, despite its preference for minimizing nuclear devastation, retaliation would be advantageous in the long run because it would make the future use of a retaliatory policy credible and so effective would be to overlook the probable lack of a relevant long run. After a nuclear exchange, future expectations, if any, would likely have very little basis in the policies of the nations prior to the exchange. Thus, even if in some cases a deterrent policy could be rationalized by an appeal to future expectations, nuclear retaliation lacks such a rationale.

Retaliation would therefore seem to be an irrational policy. If submission is preferred to retaliation, as minimizing the expected nuclear devastation one suffers, then the expression of the conditional intention to retaliate would lack credibility. The US could not expect to affect the SU's expectations about US behavior by expressing such an intention, and so the US could not decrease the probability of the SU's pursuing a strike policy by announcing its own policy of nuclear retaliation. Among sufficiently rational and informed nations, nuclear deterrence must fail. If it succeeds in the real world, then the expressed intention not to submit and to retaliate must serve, it seems, to inform the potential attacker of the would-be deterrer's real preferences, or to deceive the attacker about those preferences, or to warn the attacker to expect an irrational response to a strike policy.

But this conclusion is mistaken. We have reached it by focusing entirely on the benefits and costs of actually carrying out the conditional intention that is the core of a deterrent policy. We have failed to consider the benefits and costs of forming or adopting such a conditional intention. The argument against the rationality of nuclear retaliation, or more generally against a deterrent policy, has this structure: it is not utility maximizing to carry out the nonsubmissive, retaliatory intention; therefore it is not rational so to act; therefore it is not rational to form the intention; therefore a rational person cannot sincerely express the intention; therefore another rational and informed person cannot be deterred by the expression of the intention. The structure of the argument that I shall present and defend is: it may be utility maximizing to form the nonsubmissive, retaliatory intention; therefore it may be rational to form such an intention;

if it is rational to form the intention it is rational to act on the intention; therefore a rational person can sincerely express the intention; therefore another rational and informed person can be deterred by the expression of the intention. We shall of course have to consider why this argument succeeds and the former argument fails.

I shall therefore defend the rationality of deterrent policies and, more particularly, of nuclear retaliation. But my defense is a limited one. Indeed, among rational and informed actors, a policy of pure and simple deterrence is not rational, although it may be rational as part of a larger policy directed, among other things, at the obsolescence of deterrence. Putting my position into a historical context, I shall defend Hobbes's formulation of the first law of nature: "That every man, ought to endeavour Peace, as farre as he has hope of obtaining it; and when he cannot obtain it, that he may seek, and use, all helps, and advantages of Warre."[3] Deterrence is both an advantage of war and, among rational actors, a means to peace. Or rather, some deterrent policies may have these features. But as a means to peace, a deterrent policy looks to its own supercession. For recognition of the rationality of deterrence is inseparable from recognition of the rationality of moving, not unilaterally but mutually, beyond deterrence.

III

To give precision to our analysis of deterrence, I shall focus on situations with a very simple structure. An actor who, consistently with our previous usage, we call B, has a choice between two alternatives, y and y', where y corresponds to *advant*. If he chooses y, then another actor, A, knowing B's choice, has a choice between two alternatives, x and x', where x corresponds to *retal* and x' to *nonretal*. If B chooses y', then A may or may not have a choice between x and x' or other alternatives; initially we need suppose only that some outcome is expected. There are, then, three possible outcomes relevant to our analysis: yx, or *advant* followed by *retal*; yx', or *advant* followed by *nonretal;* and y'—, or B's choice of his alternative to *advant* followed by a possible but unspecified choice by A. Each actor orders these possible outcomes; for simplicity we assume that neither is indifferent between any two. There are then six possible orderings for each actor, and so thirty-six different possible pairs of orderings.

Only one of these thirty-six pairs determines a deterrent situation. Consider first A's orderings. Since she seeks to deter B from *advant*, she must prefer y'—, the expected outcome if B chooses his alternative action, to both yx and yx'. And since she seeks to deter B from *advant* by expressing a conditional intention to *retal* contrary to her known preferences, she must prefer yx' to yx. Now consider B's orderings. Since A seeks to deter him from *advant* by expressing her conditional intention to *retal*, he must prefer yx' to yx. If A has any need to seek to deter B from *advant*, then

3. Thomas Hobbes, *Leviathan* (London, 1651), chap. 14.

he must prefer yx' to y'—, and if she is to have any hope of deterring him, then he must prefer y'— to yx. A's ordering is: $y' - > yx' > yx$; B's ordering is: $yx' > y' - > yx$.

Let us take a brief, closer look at the outcome if B chooses y'. I shall not pursue the implications of this discussion in the present paper, although it raises issues of some interest and importance. If deterrence is to be possible, then, should B choose y', A must have a choice w (where this includes the limiting case in which she has no alternative to w) such that she prefers $y'w$ to yx' and he prefers $y'w$ to yx. If for every alternative w' such that A prefers $y'w'$ to yx', B prefers yx to $y'w'$, then, much as A might wish to deter B from choosing y she has no conditional intention sufficient. If for every alternative w'' such that B prefers $y'w''$ to yx, A prefers yx' to $y'w''$, then even though A may have a conditional intention sufficient to deter B she has no interest in using it.

Suppose then that A prefers $y'w$ to yx', and B prefers $y'w$ to yx. If B also prefers yx' to $y'w$, then A will seek to deter B from choosing y. But the expression of a conditional intention to choose x in response to y, even if fully credible, may be insufficient to deter B. For A may have an alternative w' to w such that A prefers $y'w'$ to $y'w$, but also such that B prefers yx to $y'w'$. Were B to choose y' in response to A's conditional intention to respond to y with x, then he would expect A to choose w' rather than w, so that he would be worse off than if he had ignored A's attempt to deter. However, were A to combine her expression of conditional intention to choose x in response to y with the credible expression of a conditional intention to choose w in response to y', then B, preferring $y'w$ to yx, would choose y'. In this case A is able to deter B only if she is able to combine her threat with an offer—an offer to refrain from her utility-maximizing choice in order to leave B open to her threat. Note that, although A's offer requires her not to choose her utility-maximizing response to B's choice of y', by making it she may expect an outcome $y'w$ which affords her greater utility than the outcome yx' which she would otherwise expect. Note also that B would prefer A not to be in a position to make such an offer.

It will be evident that A's conditional intention to choose a nonmaximizing w in response to y' raises precisely the same problem of rationality as her conditional intention to choose a nonmaximizing x in response to y—*retal* in response to *advant*. In both cases she must form an intention to choose a course of action in itself nonmaximizing, as part of a policy intended to maximize her expected utility. I shall not address the problem of nonmaximizing offers in this paper, but an argument for the rationality of deterrent threats can easily be applied to the offers as well.

Before proceeding to that argument let us relate our abstract treatment of deterrence to the particular issue of nuclear retaliation. In the terms in which we have posed that problem, the US corresponds to actor A, the SU to actor B. The policy of nuclear retaliation by the US corresponds

to x or *retal;* the strike policy for actor A corresponds to y or *advant.* Recall that "strike" and "retaliation" are shorthand for more complex policies; the core of a strike policy may be the threat to launch a nuclear strike should some initiative be resisted; the core of a retaliatory policy may be the refusal to acquiesce in such a threat—with, of course, the intention to retaliate should such refusal lead to a strike.

I suppose then that the US orders the possibilities: no strike > strike and no retaliation > strike and retaliation. The first preference is evident; the second preference follows from the assumption that the US wishes to minimize nuclear devastation, given that retaliation, as we have characterized it, increases its expectation of suffering such devastation. And I suppose that the SU orders the possibilities: strike and no retaliation > no strike > strike and retaliation. As I noted in the preceding paragraph, a strike policy may center on a threat; the SU's first preference need not indicate a passion for blood but only a desire to get its way by resorting to whatever threat may be needed. The SU's second preference follows from the assumption that it too wishes to minimize being the victim of nuclear devastation.

These preference orderings satisfy the requirements for a deterrent situation. I suppose that they are a plausible schematic representation of the preferences of possible real-world counterparts of the US and the SU. Thus our argument for the rationality of deterrent threats is not intended to be an enquiry into merely possible worlds. However, some of the points raised abstractly in this section should be borne in mind in any attempt to apply our argument. In particular, it is worth noting that the SU may suppose that the US has several possible responses to its no-strike policy, some of which, such as a unilateral US strike, might indeed be worse from its perspective than a strike policy coupled with US retaliation. Effective deterrence by the US may then require an offer sufficient to allay SU fears of possible unilateral US action in response to a no-strike policy. I shall not pursue this matter here, but it is essential to be aware that the components of an effective policy of nuclear deterrence are matters that require the most careful evaluation.

IV

The key to understanding deterrence, or, for that matter, the key to understanding all forms of interaction, such as agreement, that require constraints on directly maximizing behavior, is that in interaction, the probability that an individual will be in a given situation or type of situation may be affected by the beliefs of others about what that individual would do in the situation. B's willingness to put A in a situation, to face A with a choice, will be affected by his belief about how she will act in that situation, how she will choose. His belief about how she will act will be affected by his assessment of her intentions. In particular, if he knows that she is fully in control of what she does, he will, ceteris paribus, expect her to do what she conditionally intends to do should she be in that

situation. Hence the probability of A being in a given situation, insofar as her being in that situation is determined by the actions of B, is affected by A's prior intentions about what she will do in that situation.

It is of course true that, if A is rational, then her intentions must be those that it is rational for her to hold. But neither A nor B can ascertain the rationality of her intentions merely by considering the actions to which various possible intentions might commit her, and their payoffs. If B's beliefs about A's intentions partially determine what situations she will be in, then A, in forming her intentions, must consider the situations she may expect to face given the possible intentions she might form, and the payoffs from those situations. It may be tempting to suppose that it is rational to form an intention if and only if it would be utility maximizing to execute the intention. Instead we argue that it is rational to execute an intention if and only if it is utility maximizing to form it.

Let us then examine the calculations of a rational actor choosing among possible intentions. I shall restrict our analysis to the simplest case, corresponding to our analysis of deterrent situations in the preceding section. Suppose then that A must decide whether to adopt the intention to do x in a situation characterized by the performance of some action y by another actor B. Let $u(yx)$ be the utility she would expect were she to do x given y. Let x' be the alternative intention to x so that $u(yx')$ is the utility she would expect were she to act on x' given y. Let $u(y')$ be the utility she would expect were B not to do y. And let p_x be the probability that B will do y should A adopt the intention to do x given y, and $p_{x'}$ the probability that B will do y should A adopt the intention to do x' given y.

Then A's expected utility should she intend x is:

$$p_x u(yx) + (1 - p_x)u(y') .$$

And her expected utility should she intend x' is:

$$p_{x'}u(yx') + (1 - p_{x'})u(y') .$$

Our concern is with the rationality of a deterrent policy. Hence we suppose that A does not want to be faced with y, which corresponds to *advant*, so that her utility $u(y')$ is greater than both $u(yx)$ and $u(yx')$. Furthermore, we suppose that doing x, which corresponds to *retal*, is not utility maximizing for A, so that $u(yx')$ is greater than $u(yx)$. And finally, A must suppose that intending x should B do y reduces the probability of his doing y, so that $p_{x'}$ is greater than p_x.

Since A prefers facing y' to doing x' given y, and doing x' given y to doing x given y, there must be some lottery over facing y' and facing y with the intention of doing x, that A considers indifferent to the certainty of facing y with the intention of doing x'. Let p be the probability of facing y' in that lottery. Then we may express the utility of facing y with

the intention of doing x', $u(yx')$, in terms of the utilities of facing y', $u(y')$, and of facing y with the intention of doing x, $u(yx)$:

$$u(yx') = pu(y') + (1 - p)u(yx) .$$

Without loss of generality for our argument we may set $u(y') = 1$, and $u(yx) = 0$. Then:

$$u(yx') = p .$$

And so A's expected utility if she intends x given y is:

$$1 - p_x .$$

And her expected utility if she intends x' given y is:

$$p_{x'}p + (1 - p_{x'}) .$$

Suppose that A maximizes her expected utility by forming the intention to do x should B do y, that is, by forming the intention to *retal* should B *advant*. Then it must be the case that:

$$(1 - p_x) > [p_{x'}p + (1 - p_{x'})] .$$

Or equivalently:

$$[(p_{x'} - p_x)/p_{x'}] > p .$$

To interpret this condition, we note that avoiding y constitutes "deterrent success," whereas facing y and doing x constitutes "deterrent failure." Facing y and doing x' we may identify with nondeterrence. Then p is that probability of deterrent success, where the alternative is deterrent failure, that makes a deterrent policy indifferent to nondeterrence from the standpoint of the prospective deterrer. We may therefore call p the "minimum required probability" for deterrent success; it reflects the value of nondeterrence relative to deterrent success and failure. The expression $[(p_{x'} - p_x)/p_{x'}]$ is the "proportionate decrease" in the probability of being in the situation that the prospective deterrer would avoid, that is achieved by her policy of deterrence. Thus the condition states that, for a deterrent policy to be rational, the proportionate decrease that it effects in the probability of facing the undesired action, *advant*, must be greater than the minimum required probability for deterrent success.

Consider a simple example. B, a university professor in Boston, is offered a position in Dallas. His wife, A, wishes to deter him from accepting the appointment, and so tells him that, if he accepts it, she will leave him and remain in Boston, even though she would prefer to accompany

him to Dallas. Then if A is indifferent between a lottery that would offer a 70 percent chance that B would stay in Boston and a 30 percent chance that he would go alone to Dallas, and the certainty that both would go to Dallas, .7 is a minimum required probability for deterrent success. If A supposes that there is a 50 percent chance that B will accept the appointment in Dallas if she will accompany him, but only a 10 percent chance that he will accept it if she won't, then the proportionate decrease effected by deterrence in the probability that he will accept the appointment is (.5 − .1)/.5, or .8. Since .8 is greater than .7, A indeed maximizes her expected utility by her adoption of a deterrent policy, requiring her to form the conditional intention not to accompany B should he accept an appointment in Dallas.

Consider now the application of our analysis to the policy of nuclear retaliation. Deterrent success for the US lies in not facing a strike policy by the SU—a policy that intends directly, or threatens and so intends conditionally, a nuclear strike. Deterrent failure lies in being faced with such a policy and being committed to a retaliatory response—to ignoring any threat by the SU and to responding to a nuclear strike by a counterstrike. Nondeterrence lies in facing a strike policy by the SU without being committed to a retaliatory response, and so it involves acceptance of the lesser evil between acquiescing in whatever initiative the SU takes and engaging in retaliation. Given these alternatives, we may suppose that, although deterrent success is of course preferred to nondeterrence, both are strongly preferred to deterrent failure. It may indeed be better to let the Reds have their way than to be among the nuclear dead. Thus a substantial decrease in the probability of facing a strike policy by the SU is required if the deterrent policy of nuclear retaliation is to maximize the expected utility of the US and so be rational to adopt.

I shall not try to estimate the extent of this decrease or, equivalently, the minimum required probability for deterrent success. This is a difficult empirical question. What is clear is that a merely ordinal ranking of preferences over possible outcomes does not afford sufficient information to assess the rationality of a deterrent policy, either in general or in the specific case of nuclear retaliation. An actor might prefer, and strongly prefer, to avoid facing a situation brought about by some other actor doing y, but the proportionate reduction in the probability of facing y that could be effected by a deterrent policy might not be worth the expected cost of facing it with the deterrent intention. The benefits of deterrent success must always be balanced against the costs of deterrent failure, and only the relevant probabilities of being in the undesirable situation, both with and without a policy of deterrence, together with an interval measure of utility in terms of which we may calculate the minimum required probability for deterrent success, enables us to calculate the balance of benefits and costs. If our argument shows that deterrent policies in general, and nuclear retaliation in particular, may be utility maximizing, it also shows that such policies may *not* be utility maximizing, and it may

be extraordinarily difficult to determine, in a particular case, whether deterrence or nondeterrence is less disadvantageous.

But while I want to emphasize this cautionary note, I do want to insist that my argument refutes the claim that deterrence is necessarily an irrational policy because carrying out the deterrent intention is not utility maximizing. The argument for the irrationality of deterrence looks only to the costs of deterrent failure. Because there are such costs, it rejects the policy. My argument, on the other hand, relates the probability-weighted costs of deterrent failure to the probability-weighted benefits of deterrent success, in order to assess the rationality of forming the conditional, nonmaximizing intention which is the core of a deterrent policy. I claim that if it is rational to form this conditional, deterrent intention, then, should deterrence fail and the condition be realized, it is rational to act on it. The utility cost of acting on the deterrent intention enters, with appropriate probability weighting, into determining whether it is rational to form the intention. But once this is decided, the cost of acting on the intention does not enter again into determining whether, if deterrence fails, it is rational to act on it. Acting on it is part of a deterrent policy, and if expected utility is maximized by forming the conditional, deterrent intention, then deterrence is a rational policy.

V

Let us turn to some possible objections to this argument. We may forestall one counterargument by noting that, of course, if one is able to achieve the same deterrent effect by pretending to form a conditional, nonmaximizing intention as by actually forming it, then such pretense would be rational. Even if pretense offers a lesser deterrent effect, its lesser possible costs may make it rational. But there is no reason to suppose that pretense must always have as great a net benefit as the actual formation of an intention. It must be judged on the same, utility-maximizing basis as the real thing.

An objector may insist that pretense can be rational because it does not commit one to nonmaximizing behavior, but that a genuine commitment to nonmaximization cannot be rational. If it is rational to form an intention that commits one to what, ceteris paribus, would not maximize one's utility, then the utility of forming the intention must affect the utility of carrying it out, increasing it so that execution is utility maximizing. The US would, in the abstract, prefer not to engage in a nuclear exchange with the SU. Our objector admits this but urges that, if a nuclear exchange arises from a rational policy of deterrence, then the US would prefer to maintain that policy and so prefer to engage in the exchange. On his view, preference for forming a conditional intention entails preference for executing it should the condition be met.

But what reason has he for claiming this, other than his insistence on a simple, and in my view simpleminded, account of the connection between utility maximization and rationality?[4] I have shown that the adoption of an intention can be utility maximizing even though acting on it would not be, at least considered in itself. Why then should we suppose that, because adoption is utility maximizing, implementation magically becomes utility maximizing? Why should we suppose that a preference for adopting or forming an intention must carry with it a preference for implementing or executing the intention? The two preferences are logically and actually quite distinct. We may grant that in most situations one prefers to adopt an intention because one would prefer to execute it. But my argument is intended to show that this connection does not hold between conditional intentions and their implementation in deterrent situations. I have shown why the connection does not hold—because adoption of the intention affects one's expected utilities by affecting the probability that the condition for implementation will be realized.

Our objector must surely take another and stronger tack. If he allows our argument about the rationality of adopting a nonmaximizing intention, then he must claim that it may be rational to adopt an intention even though it would be, and one knows that it would be, irrational to act on it should the condition for implementing it be realized. If our objector takes this tack, then he acknowledges the rationality of some deterrent policies, but nevertheless insists that these policies, although fully rational, involve the performance of irrational actions if certain conditions are satisfied. How then does his position differ from mine, in which I claim that deterrent policies may be rational, and if rational, involve the performance of actions which, in themselves and apart from the context of deterrence, would be irrational, but which, in that context, result from rational intentions and so are rational?[5] Surely he grants the substance of my argument but expresses his agreement in a misleading and even paradoxical way, insisting that actions necessary to a rational policy may themselves be irrational. To assess an action as irrational is, in my view, to claim that it should not be, or have been, performed. If our objector accepts deterrent policies, then he cannot consistently reject the actions they require and so cannot claim that such actions should not be performed.

Suppose, then, that our objector confronts my position head on and rejects the rationality of deterrent policies. He insists that the execution of an intention must take precedence, rationally, over its adoption. He

4. If preference is necessarily revealed in behavior, then choosing a nuclear exchange shows that one prefers it to one's alternatives. Conceptually, we can (and many economists and game theorists do) fit preference and choice so tightly together that nothing could count as non-utility-maximizing behavior. But this mode of conceptualization is a Procrustean bed for the treatment of such issues as the rationality of deterrence.

5. How his position may differ is made clear by David Lewis, "Devil's Bargains and the Real World," in *The Security Gamble: Deterrence Dilemmas in the Nuclear Age*, ed. Douglas MacLean (Totowa, N.J.: Rowman & Allanheld, in press). I begin a rejoinder to Lewis in "Response to the Paradox of Deterrence," in MacLean, ed.

must insist that it is rational to form an intention if and only if one maximizes one's expected utility both in forming it and in executing it. If either condition fails, then formation of the intention is not rational.

This objector insists that the rationality of an action is always to be assessed *from now*, in the words of Bernard Williams.[6] The rationality of an action is to be assessed from the point at which the question, not of intending it, but of performing it, arises. And this is, I think, the heart of the matter. In taking this position the objector applies the utility-maximizing standard of rationality in the way generally approved by economists, decision theorists, and game theorists. But he, and they, are mistaken. The 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.

A utility-maximizing policy may include non-utility-maximizing performance. Deterrence exemplifies this. The expected utility of a policy is the sum of the probability-weighted expected utilities of the performances it allows or requires. The apparent paradox, that a utility-maximizing policy may contain non-utility-maximizing performances, is resolved in the realization that altering the performances need not be independent of altering their probabilities. An assessment that begins and remains at the level of the performances neglects this crucial fact. And so the actor who assesses the rationality of his actions only from now, from the point at which the question of performance arises, may expect a lesser overall utility than the actor who assesses the rationality of her actions in the context of policies, who adjusts performances so that the probability-weighted sum of their utilities is greatest.

Our objector will say that the policy maximizer allows her choices to be ruled by the dead hand of the past, whereas he, the performance maximizer, lives and chooses in the present. But our objector is mistaken. Unable to escape the burden of choice, the performance maximizer must, choosing in the present, keep in mind that his attempt to maximize utility in the present performance is constrained by his future attempts to maximize utility on the occasion of each successive performance. He is ruled by the unborn, and perhaps never-to-be-born, hands of his possible futures. And his yoke is the worse. Maximization is the policy maximizer's goal, but the performance maximizer's fate.[7]

Before leaving our objector to that fate, let us note carefully that the reply to him does not insist that one should maximize in the long run rather than the short run. The would-be deterrer who fails to deter and who must then make good on her threat in order to carry out her

6. Bernard Williams, *Moral Luck* (Cambridge: Cambridge University Press, 1981), p. 35.

7. I expand on this point in "Response to the Paradox of Deterrence."

conditional intention, is not maximizing at all. Her reason for sticking to her guns is not to teach others by example, not to improve her prospects for successful deterrence in the future, or anything of the sort. Her reason is simply that the expected utility or payoff of her failed policy depended on her willingness to stick to her guns.

Let us suppose that each person or nation—each actor—knew (never mind how!) that but once in his life he would be in a situation in which, by convincing another actor that he would respond in a nonmaximizing way to a possible choice of the other, he could increase his expected utility by reducing the probability that the other would make that choice. Here, if the other is not deterred, carrying out the nonmaximizing response can, *ex hypothesi*, have no effect on the actor's credibility or on future deterrence. Yet he can hope to deter only if the other believes that he will, or at least may, make that nonmaximizing response. And adopting a genuine policy of deterrence may be the only way of bringing about that belief, or increasing its strength, in the other person. Even in this one-shot situation, a deterrent policy, committing one to a nonmaximizing choice should deterrence fail, may be utility maximizing. If I have convinced you of this, then I have accomplished my most important task in this essay, because only those convinced can have a proper understanding, not only of deterrence, but also of the whole range of situations, including most prominently generalized Prisoners' Dilemmas, in which policies that require nonmaximizing behavior are utility maximizing, and so rational.[8] And what these policies effect is throughout the same—to alter the probabilities of an actor's being in certain situations, facing certain choices. Only in understanding this do we begin to appreciate the true characteristics and complexity of utility-maximizing rationality.

VI

I have referred in passing to the expression of a conditional intention to *retal* as a threat. And the argument that I have advanced for the rationality of a deterrent policy is indeed an argument for the rationality of threat enforcement. If the expected gain from deterrence exceeds the expected cost of carrying out the deterrent threat, where each expectation is probability weighted, and if no less costly means of deterrence is available, then the rational actor sincerely threatens and enforces her threat should it fail to deter.

Not all threats, we may pause to note, are properly deterrent. The kidnapper threatens the parents of his victim with the death of their child should they fail to pay; it would be perverse to say that he seeks to deter them from nonpayment. But I shall not attempt an analysis of threats here. My purpose in introducing the conception of threat is to

8. I discuss this, although obscurely, in "Reason and Maximization," *Canadian Journal of Philosophy* 4 (1975): 427–30. Matters should be clearer in my *Morals by Agreement* (Oxford: Oxford University Press, in press), chap. 6.

broaden the perspective of our analysis so that it embraces both threatener and threatened, and in this perspective we shall find a new and problematic dimension in our argument.

If we think of nuclear retaliation as a policy of threat enforcement, yet we must note immediately that it is also a policy of threat resistance. The US threatens nuclear retaliation to deter a strike by the SU, but a strike policy, as we have described it, may center on the issuance of a credible threat of nuclear attack should some initiative be opposed, and retaliation thus embraces resistance to such a threat. In the context of nuclear deterrence each party may be viewed both as threatener and as threatened, both as a potential threat enforcer and as a potential threat resister. Not all threat situations involve this symmetry, but the standpoints of threatener and threatened are themselves significantly parallel. For each must decide whether to adopt an intention—to enforce a threat or to resist a threat. The enforcer seeks to avoid that situation in which enforcement would be required; the resister seeks to prevent that situation in which resistance would be required. The argument of Section IV may be adapted to show the rationale for both threat enforcement and threat resistance. Since, taken together, enforcement and resistance make threat behavior unprofitable, the existence of parallel rationales may cast doubt on the rationality of any policy involving threats, and so on a policy of deterrence.

Let us consider briefly how the argument of Section IV applies to enforcement and resistance. Both the would-be threat enforcer and the would-be threat resister seek to reduce the probability of being in an undesirable situation (having one's threat ignored/facing a credible threat) by expressing a conditional intention to respond in a mutually costly way in that situation. Enforcement/resistance success lies in avoiding the undesirable situation; enforcement/resistance failure lies in having to carry out one's conditional intention. The minimum required probability for enforcement/resistance success is defined as the probability of that success in the lottery between success and failure that the enforcer/resister considers indifferent to no enforcement/no resistance. A policy of threat enforcement/threat resistance is rational only if the proportionate decrease that it effects in the probability of having one's threat ignored/facing a credible threat is greater than the minimum required probability for enforcement/resistance success.

The parallel rationales that can be constructed for threat enforcement and threat resistance may seem to show the overall irrationality of threat behavior. For if both enforcement and resistance are rational, then either the worst case prevails, in which a threat is issued, ignored, and executed, or the prethreat situation prevails, no threat being issued since, if it were, it would be ignored and then executed. But although there is a deep irrationality in threat behavior, the parallel rationales do not themselves suffice to demonstrate it. For they show only that the structure of the argument for enforcement is the same as that for resistance. They do

not show that, in a given situation, threat enforcement and threat resistance are equally rational or irrational.

We may illustrate this by our core example—nuclear deterrence. Suppose that the SU were to announce a policy of deterrence-resistance. It will carry out, or threaten, a nuclear strike if it considers that a retaliatory response would be costly to the US—if it believes that the maximizing US response would be acquiescence or submission.

As we noted in Section III, the SU prefers strike and no retaliation to no strike, and no strike to strike and retaliation. A policy of deterrence-resistance is rational for the SU only if the proportionate decrease that it effects in the probability of a US policy of retaliation is greater than the minimum required probability for the success of deterrence-resistance. But this is the probability of strike and no retaliation in that lottery between strike and no retaliation and strike and retaliation that the SU finds indifferent to the certainty of no strike. No strike represents, in effect, acceptance of the status quo; we may plausibly suppose that the SU would require a very high probability of gain—of the US acquiescence entailed in strike and no retaliation—and a correspondingly low probability of loss—of the nuclear exchange entailed in strike and retaliation—before it would be indifferent between such a lottery and the status quo. We may plausibly suppose that deterrence-resistance will not seem to the SU to be a utility-maximizing policy.

The US, as we also noted in Section III, prefers no strike to strike and no retaliation, and strike and no retaliation to strike and retaliation. Thus as we established in Section IV, deterrence is a rational policy for the US only if the proportionate decrease that it effects in the probability of a strike policy by the SU is greater than the probability of no strike in the lottery between no strike and strike and retaliation that the US finds indifferent to the certainty of strike and no retaliation. Although we have refrained from attempting to estimate this probability, except to suggest that it is likely to be high, yet we may note that strike and no retaliation represents, not the status quo, but a real worsening of the situation of the US. Even though a nuclear exchange is a greater worsening, yet we may plausibly suppose that the US would not require a very high probability of maintaining the status quo implicit in no strike, and a very low corresponding probability of loss through nuclear exchange, to be indifferent between such a lottery and the loss implicit in no retaliation. Although any firm judgment must be beyond armchair competence, it may well be the case that nuclear retaliation is a rational policy for the US, although resistance to deterrence is not a rational policy for the SU.

Thus the parallel between the rationales for threat enforcement and threat resistance does not in itself show the irrationality of a policy of deterrence. However, even if threat behavior is rationally justifiable from the standpoint of a particular actor, there is a need for mutually agreed measures to remove the threat-inviting context. Fundamental to Hobbes's analysis of the state of nature is the need to exit through the acceptance

worsen the condition of those against whom it is directed. To resort to such a policy is to reject the prospect of cooperative interaction with others.

Nuclear retaliation, as a deterrent policy, is directed at protecting the retaliator from being victimized by any actor willing to engage in a first strike. It is, then, not to seek to redistribute benefits in a way more favorable to the would-be deterrer than could be expected in the absence of interaction but, rather, to ensure that her situation is not worsened in terms of that baseline. It is directed at upholding, rather than subverting, the requirement that human society be a cooperative venture for mutual advantage.

In itself, of course, nothing could be less cooperative, less directed at mutual advantage, than the use of nuclear weapons. But a retaliatory, deterrent policy is directed at preventing such use—directed at maintaining those conditions in which societies may be brought to recognize the benefits of cooperation. A policy of nuclear deterrence clearly has failed if a nuclear exchange occurs. But the serious alternative to such a policy, in the absence of agreement to eschew all threat behavior, can only be the willingness to accept victimization, to suffer passively a nuclear strike or to acquiesce in whatever the potential striker demands as the price of its avoidance.

Morality, in my view, follows rationality. Practical rationality is concerned with the maximization of benefit; the primary requirements of morality are that in maximizing benefit, advantage must not be taken and need not be given.[11] Nuclear deterrence, despite its horrific character, is then a moral policy—a policy aimed at encouraging the conditions under which morally acceptable and rational interaction among nations may occur. If we agree that the idea of society as a cooperative venture for mutual advantage, and the related proviso against benefiting through interaction that worsens the condition of others, express a fundamental moral ideal, then the willingness to maintain those conditions under which this ideal may be realized, and the refusal to acquiesce in measures that would subvert it, must themselves be the objects of moral approval rather than censure.

Rational nations, recognizing the need to seek peace and follow it given the costs of war, can unilaterally renounce the first use of nuclear weapons and thereby end all strike policies. Rational nations can mutually agree to destroy their holdings of nuclear weapons, at least insofar as these weapons are directed against each other, and so can end all deterrent policies. Since the knowledge that brought nuclear weapons into being will not disappear, we cannot expect a world fully free of nuclear threats. We can only minimize a peril that cannot be exorcised. But to understand the conditions under which we may rationally agree to the mutual abandonment of deterrent and other threat policies, we must first understand the rationale of deterrent policies and the role of these policies in maintaining the conditions of acceptable international interaction. Hobbes conjoins two fundamental requirements in relating the law and the right of nature: "To seek Peace, and follow it" and "by all means we can, to defend our selves."[12] Hobbes understands that these requirements are mutually supportive; a correct understanding of nuclear deterrence supports his view.

11. Neither utilitarians nor Kantians will find this conception of morality to their taste. I cannot defend it here, but see Gauthier, *Morals by Agreement*.

12. Hobbes, *Leviathan*, chap. 14.