

OPTIMAL DETERRENCE

Steven J. Brams
Department of Politics
New York University
New York, NY 10003

D. Marc Kilgour
Department of Systems Design Engineering
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1

ABSTRACT

Deterrence means threatening to retaliate against an attack in order to deter it in the first place. It is widely held that only through continuing mutual deterrence has a nuclear confrontation of the superpowers been avoided. Yet the central problem with a policy of deterrence is that the threat of retaliation may not be credible if retaliation leads to a worse outcome - perhaps a nuclear holocaust - than a side would suffer from absorbing a limited first strike and not retaliating.

The optimality of mutual deterrence is analyzed by means of a "deterrence game," in which each player chooses a probability (or level) of preemption, and of retaliation if preempted. The Nash equilibria, or stable outcomes, in this game duplicate those in the game of Chicken, on which it is based, except for a "deterrence equilibrium," at which the players never preempt but are always prepared to retaliate with a probability above a calculable threshold. This equilibrium is Pareto-superior, dynamically stable, and -- when supported by "robust threats" -- as invulnerable as possible to misperceptions or miscalculations by the players. How robust the deterrence policies of the superpowers are today is unclear, though their precommitments to retaliate if attacked are consistent with probabilistic threats in the deterrence game.

1. Introduction

The policy of deterrence, at least to avert nuclear war between the superpowers, has been a controversial one. The main controversy arises from the threat of each side to visit destruction on the other in response to an initial attack. This threat would seem irrational if carrying it out would lead to a nuclear holocaust -- the worst outcome for both sides. Instead, it would seem better for the side attacked to suffer some destruction rather than retaliate in kind and, in the process of devastating the other side, seal its own doom in an all-out nuclear exchange.

Yet, the superpowers persist in their adherence to deterrence, by which we mean a policy of threatening to retaliate to an attack by the other side in order to deter such an attack in the first place. To be sure, nuclear doctrine for implementing deterrence has evolved over the years, with such appellations as "massive retaliation," "flexible response," "mutual assured destruction" (MAD), and "counterforce" giving some flavor of changes in U.S. strategic thinking.

All such doctrines, however, entail some kind of response to a Soviet nuclear attack. They are operationalized in terms of preselected targets to be hit, depending on the perceived nature and magnitude of the attack. Thus, whether U.S. strategic policy at any time stresses a retaliatory attack on cities and industrial centers (countervalue) or on weapons systems and armed forces (counterforce), a certain response of some kind to an attack is not the issue. The issue rather is what kind of threatened response, or second strike in the parlance of deterrence theory, is most efficacious in deterring an initial attack, or first strike.

This is the issue we address in this paper, though not in the usual way. Instead of trying to evaluate the relative merits of concrete nuclear

retaliatory doctrines, we shall define these doctrines somewhat more abstractly in terms of "probabilistic threats." More specifically, by letting threats vary along a single continuous dimension from certain retaliation to no retaliation, we can compare different levels of threats in terms of the expected payoffs that they yield in a game. Additionally, by introducing probabilities of a first strike (or preemption) by both sides, we can analyze the relationship between preemption and retaliation probabilities and game outcomes.

Because the expected payoffs of probabilistic preemption and retaliation have certain equivalents, in that a player would be indifferent between choosing a lottery (over nonpreemption/preemption or nonretaliation/retaliation) and a sure thing (a reduced level of preemption or retaliation), they can be interpreted in terms of levels of preemption and retaliation short of full-fledged first and second strikes. The first question we seek to answer is what levels render certain outcomes stable, in a sense to be specified later.

In the game we use to model deterrence, which is derived from the game of Chicken but is not Chicken itself, we identify four stable outcomes, or equilibria, three of which correspond to those in Chicken. The new equilibrium which emerges when we incorporate the possibility of (probabilistic) preemption and retaliation into Chicken we call the "deterrence equilibrium." It corresponds to the cooperative outcome in Chicken (never preempt), which by itself is unstable; in the new (deterrence) game, this outcome is rendered stable by the threat of retaliation above a calculable threshold, which makes preemption irrational.

But a threshold alone does not specify what level of threat (above

this threshold) is optimal. Accordingly, we suggest a theoretical calculation of "robust threats" that makes retaliatory threats as invulnerable as possible to misperceptions or miscalculations by the players. We also indicate how precommitments to carry out these threats are in fact made credible, at least on a probabilistic basis, by the superpowers.

We think the deterrence equilibrium -- and robust threats that support it -- is superior to any other equilibrium in the game we postulate as a model of deterrence. To be sure, this equilibrium is imperfect in the sense that it is irrational to carry out one's threats; however, because it renders preemption irrational, even when one thinks one's opponent might preempt, it is hard to see why retaliation would ever be necessary, at least in theory. This theoretical rationale for a particular kind of deterrence, coupling a no-first-use policy with robust threats, appears to us the best one can do in a world that seems to make superpower confrontations unavoidable.

The challenge facing the policy maker is to prevent such confrontations from escalating into nuclear war. As deleterious as threats are to the development of trust and good will, we conclude that they are inescapable for deterrence to be effective. It is far less clear whether the threats the superpowers hurl at each other today, and their concomitant actions to indicate the threats are not empty, are at an optimal level.

2. Deterrence and the Game of Chicken

There is a large literature on deterrence, but little of it is explicitly game-theoretic. That which is, or is pertinent to game-theoretic formulations, is discussed by Brams from both a theoretical

and empirical perspective, so we shall not review it here.¹ Suffice it to say that we believe game theory not only provides a framework uniquely suited to capturing the interdependent strategic calculations of players but also that it can be adapted to modelling the threats necessary to deter an opponent from taking untoward action against oneself.

To incorporate threats into the structure of a game, we shall assume that players can precommit themselves to carrying out their threats with a given probability. Exactly how they do so will be considered later, but for now we shall assume precommitments are allowed by the rules of the game.

Because a game is defined by the rules that describe it, there is no problem in permitting precommitments as long as they are not inconsistent with other rules. In fact, as we shall show, the major issue precommitments raise is the rationality of holding to them in the play of a game. We shall discuss this issue after deriving the equilibria of the so-called Deterrence Game and analyzing their properties.

The Deterrence Game is based on the two-person game of Chicken, which we shall define and analyze in this section. In Chicken, each player can choose between two strategies: cooperate (C) and not cooperate (\bar{C}), which in the context of deterrence may be thought of as "not attack" and "attack", respectively. These strategies lead to four possible outcomes, which the players are assumed to rank from best (4) to worst (1). These rankings are shown as ordered pairs in

Figure 1

the outcome matrix of Figure 1, with the first number indicating the rank assigned by the row player, and the second number indicating the rank assigned by the column player. Chicken is defined by the following

FIGURE 1
OUTCOME MATRIX OF CHICKEN

		Cooperate (C)	Do not cooperate (\bar{C})
		Row	Cooperate (C) (3,3) Compromise
Do not cooperate (\bar{C})	(4,2) Row "wins", Column "loses"	(1,1) Disaster	

Key: (x,y) = (rank of Row, rank of Column)

4 = best; 3 = next best; 2 = next worst; 1 = worst

Circled outcomes are Nash equilibria

outcome rankings of the strategy combinations of the two players:

1. Both players cooperate (CC) -- next-best outcome for both players: (3,3).
2. One player cooperates and the other does not ($\bar{C}C$ and $C\bar{C}$) -- best outcome for the player who does not cooperate and next-worst outcome for the player who does: (2,4) and (4,2).
3. Both players do not cooperate ($\bar{C}\bar{C}$) -- worst outcome for both players: (1,1).

Outcomes (2,4) and (4,2) in Figure 1 are circled to indicate that they are Nash equilibria: neither player (Row or Column) would have an incentive to depart from these outcomes because he would do worse if he did. For example, from (2,4) Row would do worse if he moved to (1,1), and Column would do worse if he moved to (3,3). By contrast, from (3,3) Row would do better if he moved to (4,2), and Column would do better if he moved to (2,4).

There is a third Nash equilibrium in Chicken, but it is not in pure strategies, or specific strategies that players would choose with certainty. Rather, it is in mixed strategies, which are defined by a probability distribution over a player's pure strategies. Because the calculation of equilibria involving mixed strategies requires that payoffs be given in cardinal utilities -- not just ordinal ranks -- we will postpone discussion of these strategies and the third equilibrium until the development of the Deterrence Game in section 3.

The shorthand verbal descriptions given for each outcome in Figure 1 suggest the vexing problem the players confront in choosing between C and \bar{C} : by choosing \bar{C} , each can "win" but risks disaster; by choosing C, each can benefit from compromise, but also can "lose." Each of the Nash equilibria shown in Figure 1 favors one player over

the other, and the stability of these equilibria as such says nothing about which of the two -- if either -- will be chosen.

Other concepts of equilibrium distinguish (3,3) as the unique stable outcome, but the rules of play that render compromise stable presume that the players (1) act nonmyopically or farsightedly and (2) cannot threaten each other.² If threats are possible in repeated play of Chicken under still different rules, the stability of (3,3) is undermined.³

The effect that threats may have in Chicken is not hard to grasp. If one player (say, Row) threatens the other player (Column) with the choice of \bar{C} , and this threat is regarded as credible, Column's best response is C, leading to (4,2).

Clearly, the player with the credible threat -- if there is one -- can force the other player to back down in order to avoid (1,1). Although Row would "win" in this case by getting his best outcome, Column would not "lose" in the usual sense by getting his worst outcome, but instead his next-worst.

This is because Chicken is not a constant-sum game, in which what one player wins the other player loses. That is why we have put "win" and "lose" in quotation marks here and in Figure 1. In nonconstant-sum games like Chicken, the sum of the players' payoffs at each outcome (if measured cardinally by utilities rather than ordinally by ranks) is not constant but variable. This means that both players may do better at some outcomes [e.g., (3,3)] than others [e.g., (1,1)]. Outcomes, such as (1,1) in Chicken, which are inferior for both players to some other outcomes in a game, are called Pareto-inferior; those outcomes which are not Pareto-inferior are Pareto-superior, as are

the other three outcomes in Chicken.

We have shown that Chicken is vulnerable to the use of threats, by which we mean precommitment (before the play of the game) to the choice of a strategy by one player in order to force the other player to choose a strategy, and hence an outcome (defined by a pair of strategy choices), favorable to the threatener. For a threat to be effective (i.e., force the threatened player to choose the strategy the threatener prefers), it must be credible -- the threatened player must believe that the threatener will in fact carry out his threat.

Thus, for example, if Column did not believe that Row would actually choose \bar{C} in Chicken (e.g., because he himself also threatened to choose \bar{C}), Column presumably would choose \bar{C} in the belief that Row would back down and choose \bar{C} , leading to Column's best outcome of (2,4). Of course, if Column's belief were mistaken, the outcome for both players would be disastrous. In the Deterrence Game, we shall explore how mutual threats in Chicken may induce compromise rather than push the players toward the precipice.

Chicken is not the only game vulnerable to threats. There are 78 distinct strict ordinal 2×2 games, in which two players, each with two strategies, can strictly rank the four outcomes from best to worst. In 46 of them, one or both players has "threat power" of either a "compellent" or "deterrent" kind.⁴ Chicken, however, is the only one of the 78 games that satisfies the following four conditions:

1. Symmetry: the players rank the outcomes along the main diagonal (CC and $\bar{C}\bar{C}$) the same; their rankings of the off-diagonal outcomes ($C\bar{C}$ and $\bar{C}C$) are mirror images of each other.
2. Cooperation is preferable to noncooperation: both players prefer CC to $\bar{C}\bar{C}$.

3. Unilateral noncooperation helps the noncooperator and hurts the cooperator: Row prefers $\bar{C}C$ to CC to $C\bar{C}$, and Column prefers $C\bar{C}$ to CC to $\bar{C}C$.
4. Retaliation for noncooperation is irrational: if one player does not cooperate (i.e., the initial outcome is $\bar{C}C$ or $C\bar{C}$), retaliation by the other player (to $\bar{C}C$) is worse for the retaliator (as well as the player whom he retaliates against).

It is evident that all except condition 1, which we assume so as to pose the same strategic dilemma for each player, conspire to make Chicken a harrowing game to play. Cooperation is at the same time desirable (condition 2) and undesirable (condition 3). But the crux of the dilemma is that if one player is intransigent (i.e., noncooperative), the other player has good reason not to be (condition 4). If condition 4 does not obtain, but instead $\bar{C}C$ is better than $\bar{C}C$ and $C\bar{C}$ for the cooperative player, then the resulting game is Prisoners' Dilemma, which presents the players with a very different kind of strategic problem.

We believe, however, that the heart of the problem with deterrence, especially of the nuclear kind, is the apparent irrationality of retaliating against a first strike by an opponent.⁵ What sort of threats (if any) are credible and will deter a first strike, so as not to put one in the unenviable position of having to decide whether to retaliate and court mutual annihilation? When is a policy of deterrence involving mutual threats of retaliation stable? How can players make their precommitments to retaliate compelling? We shall explore these and other questions in our analysis of the Deterrence Game, which permits the players to choose both levels of preemption and levels of retaliation.

3. The Deterrence Game (with Preemption and Retaliation Probabilities)

The Deterrence Game is defined by the following rules:

1. The final outcome will be one of the four outcomes of Chicken. The payoffs are the same as those of Chicken, except that cardinal utilities replace ordinal rankings. Thus r_4 and c_4 signify the highest payoffs for Row and Column, respectively, r_1 and c_1 the lowest, etc.
2. The players do not choose initially between C and \bar{C} , as in Chicken, but instead choose (unspecified) actions that have associated a nonpreemption probability (s for Row and t for Column) and a complementary preemption probability ($1-s$ for Row and $1-t$ for Column). With these probabilities, the actions will be interpreted as cooperative (C) and noncooperative (\bar{C}) strategy choices, respectively.
3. If both players' initial choices are perceived as the same, the game ends at that position (i.e. CC or $\bar{C}\bar{C}$). If one player's choice is perceived as C and the other's as \bar{C} , the former player then chooses subsequent actions with an associated nonretaliation probability (p for Column and q for Row) and a complementary retaliation probability ($1-p$ for Column and $1-q$ for Row). With the retaliation probability, the conflict is escalated to the final outcome $\bar{C}\bar{C}$; otherwise it remains (at CC or $\bar{C}\bar{C}$).
4. The players choose their preemption probabilities and retaliation probabilities before play of the game. Play commences when each player simultaneously chooses initial actions that may be interpreted as either C or \bar{C} , with associated

preemption probabilities. One player may then choose subsequent actions, according to rule 3, with the associated retaliation probability specified at the beginning of play.

Figure 2 The Deterrence Game is represented in Figure 2. Note that besides the fact that the initial strategy choices of the two players are probabilities (with assumed underlying actions), rather than actions (C and \bar{C}) themselves, this payoff matrix differs from the Figure 1 outcome matrix in having expected payoffs rather than (certain) payoffs in its off-diagonal entries. This is because we assume that if one player is perceived to preempt, the other player's (probabilistic) retaliation will be virtually instantaneous, so it is proper to include in the off-diagonal entries a combination of payoffs -- reflecting both possible retaliation and possible nonretaliation -- by means of an expected value.

We assume, of course, that $0 \leq s, t, p, q \leq 1$ because they represent probabilities. To simplify subsequent calculations, we normalize the payoffs of the players so that the best and worst payoffs are 1 and 0, respectively. Hence,

$$0 = r_1 < r_2 < r_3 < r_4 = 1$$

$$0 = c_1 < c_2 < c_3 < c_4 = 1$$

Because we assume the preemption and retaliation probabilities are chosen independently by the players, the expected payoffs for Row and Column are simply the sums of the four payoffs (expected payoffs) in the Figure 2 matrix, each multiplied by the probability of its occurrence:

$$E_R(s, q; t, p) = str_3 + (1-s)tp + s(1-t)qr_2;$$

$$E_C(t, p; s, q) = stc_3 + s(1-t)q + (1-s)tpc_2.$$

In the Appendix we show that there are effectively four Nash equilibria

FIGURE 2
REPRESENTATION OF DETERRENCE GAME

		Column	
		t	1-t
Row	s	(r_3, c_3)	$q(r_2, c_4) + (1-q)(r_1, c_1)$ $= (qr_2, q)$
	1-s	$p(r_4, c_2) + (1-p)(r_1, c_1)$ $= (p, pc_2)$	$(r_1, c_1) = (0, 0)$

Key: $(r_i, c_j) = (\text{payoff to Row}, \text{payoff to Column})$

$r_4, c_4 = \text{best}$; $r_3, c_3 = \text{next-best}$; $r_2, c_2 = \text{next-worst}$; $r_1, c_1 = \text{worst}$
 $s, t = \text{probabilities of nonpreemption}$; $p, q = \text{probabilities of nonretaliation}$

Normalization: $0 = r_1 < r_2 < r_3 < r_4 = 1$; $0 = c_1 < c_2 < c_3 < c_4 = 1$

in the Deterrence Game, and that they can be grouped into three classes:

I. Deterrence Equilibrium: $s = 1, q \leq c_3; t = 1, p \leq r_3$. This equilibrium is one in which the players never preempt ($s = t = 1$), but Row retaliates with probability $1-p > r_3$ and Column retaliates with probability $1-q > c_3$. Essentially, these inequalities ensure that a player's expected payoff as the sole preemptor -- p for Row and q for Column, as shown in the off-diagonal entries in Figure 2 -- is not greater than what is obtained from the cooperative outcome of the underlying Chicken game, with payoffs (r_3, c_3) .

II. Preemption Equilibria: (1) $s = 1, q = 0; t = 0, p$ arbitrary; (2) $s = 0, q$ arbitrary; $t = 1, p = 0$. The first equilibrium is certain preemption by Column and no retaliation by Row; because Row is deterred by Column's initiative, Column's retaliation probability is arbitrary since it never comes into play. The second equilibrium is analogous, with the roles of Column and Row switched. At these equilibria, the outcomes of the Deterrence Game are the outcomes of the underlying Chicken game associated with wins for Column and Row (discussed in section 2), with payoffs $(r_2, 1)$ and $(1, c_2)$, respectively.

III. Naive Equilibrium:

$$s = \frac{c_2}{1-c_3+c_2}, q = 1; t = \frac{r_2}{1-r_3+r_2}, p = 1.$$

At this equilibrium, each player preempts with some nonzero probability, (which depends on the other player's payoffs and is always less than one) but never retaliates. Each of these preemption probabilities in fact (see Appendix) makes the opponent indifferent as to his level of preemption; in other words, a player's expected payoff depends only

on his opponent's, and not his own, level of preemption. Because retaliation would only degrade these expected payoffs, it is suboptimal. As shown in the Appendix, however, the Naive Equilibrium is Pareto-inferior to the Deterrence Equilibrium, which is the reason for our nomenclature. It corresponds to the mixed-strategy equilibrium of the underlying Chicken game (discussed but not given in section 2), which is similarly deficient as well as difficult to interpret as a one-shot choice of rational players in this game.

4. Rational Play in the Deterrence Game

Of the four Nash equilibria, only the Deterrence Equilibrium in class I depends on the possibility of retaliation -- specifically, pre-committed threats to respond (at least probabilistically) to a provocation when it is viewed as equivalent to the choice of \bar{C} . Such threats distinguish the Deterrence Game from the underlying game of Chicken, in which retaliation against the choice of \bar{C} is not permitted.

Note that the two Preemption Equilibria in class II, and the one Naive Equilibrium in class III, occur only when retaliatory threats are never used ($p = 1$ or $q = 1$ or both). They correspond precisely to the three Nash equilibria in Chicken and so introduce no new element into the analysis of deterrence beyond what was earlier provided by Chicken. However, when a threat structure is added to Chicken to give the Deterrence Game, a qualitatively different equilibrium (the Deterrence Equilibrium) emerges in the latter game that demonstrates how threats can work to the advantage of both players to stabilize the Pareto-superior cooperative outcome (r_3, c_3) , which is unstable in Chicken without the possibility of retaliation.

Because the Deterrence Equilibrium depends fundamentally on threats,

it is not surprising that it is neither perfect nor subgame-perfect in the sense of Selten.⁶ Nevertheless, the Deterrence Equilibrium possesses a dynamic-stability property that should, once the equilibrium forms, contribute to its persistence in repeated play. That is to say, given the players are at the Deterrence Equilibrium, if one player (say Column) for any reason suspects that the other player (Row) may contemplate preemption, thereby rendering $s < 1$, Row can do no better than continue to choose $t = 1$.

In other words, even should Row think he might be preempted, he should still continue to refuse to preempt, in order to keep his expected payoff at its maximum. This obviates the problem that Schelling called "the reciprocal fear of surprise attack" that leads inexorably to preemption.⁷

We prove this dynamic-stability property of the Deterrence Equilibrium in the Appendix, which shows in effect that any perceived departures of s or t from 1 will not initiate an escalatory process whereby the players are motivated to move closer and closer toward preemption for certain. The fact that the Deterrence Equilibrium is impervious to perturbations in s or t means that the players, instead of being induced to move up the escalation ladder, will have an incentive to move down should one player deviate from $s = t = 1$.

The restoration of the Deterrence Equilibrium depends on probabilistic threats of retaliation that satisfy

$$0 < q < c_3, 0 < p < r_3. \quad (1)$$

But note that if deterrence for any reason should fail, it is irrational to retaliate, even on a probabilistic basis, because retaliation leads to a worse outcome for the threatener, having to carry out his threat,

as well as for the player who preempted and thereby provoked retaliation.

The apparent irrationality of retaliating in the Deterrence Game is, as we indicated earlier, precisely what makes the Deterrence Equilibrium imperfect. Despite its imperfectness, we believe there are at least two ways in which it may be strengthened, one theoretical and one practical.

In theory, all threats which satisfy inequalities (1), given that $s = t = 1$, define a Deterrence Equilibrium. But in the intervals defined by (1), which values of p and q should be used? One of us proposed, as most insensitive to misperceptions or miscalculations by the players, robust threats,

$$q = \frac{c_3 - c_2 r_3}{1 - c_2 r_2}, \quad p = \frac{r_3 - r_3 c_3}{1 - c_2 r_2},$$

which are easily shown to satisfy (1).⁸ Such threats, when carried out, are equally damaging to the preemptor, and equally costly to the retaliator, whichever strategy either perceives the other might choose at the start (preempt or not). This property makes each player's preemption decision independent of his reading of his opponent's choice -- the damage or cost will be the same whatever he chooses -- and should serve to enhance the stability of the Deterrence Equilibrium.

A by-product of robust threats is that they render nonpreemption (strategies s and t in the Figure 2 Deterrence Game) strictly dominant -- better for each player whatever his opponent does -- and hence unconditionally best. This, of course, is not true of the C strategies in Chicken, which are undominated -- sometimes best (when the opponent chooses \bar{C}) and sometimes not (when he chooses C).

In practice, the Deterrence Equilibrium depends on the credibility

of threats satisfying (1). But how does a player persuade his opponent that he will retaliate if attacked, even though retaliation would be irrational at the time it is undertaken?

In the case of the superpowers, both the United States and the Soviet Union have institutionalized detailed procedures for responding to a nuclear attack that are designed to ensure -- insofar as possible -- that retaliation will occur, even if communication, command, control, and intelligence (C³I) capabilities are damaged by the attack.⁹ However, although each side promises that a first strike will inevitably be met by a second strike, there is significant uncertainty about each side's likely response

because of a number of operational factors, including problems related to identifying the attacker, identifying the magnitude of the attack, failures of weapons being used for the first time on a massive scale, problems of communication and control, lack of resolve, and the like. In light of these difficulties, both sides have, not surprisingly, resorted less to making probabilistic threats and more to employing their certain equivalents -- usually controlled steps up the escalation ladder.

These . . . may be thought of as probabilistic threats insofar as they give an opponent a better idea of how close each side is moving toward full-scale retaliation -- that is, they indicate more palpably the probability that the opponent will carry out a threat and what its expected damage will be.

So far, fortunately, these probabilistic threats have been sufficient to persuade the two sides to back off, beyond a certain point, from continued escalation.¹⁰

We conclude that: (1) the deterministic threats proclaimed by the superpowers today are, in truth, probabilistic (as we have modeled them); and (2) they have in fact deterred nuclear war. Moreover, there seems little doubt that both sides have precommitted themselves to retaliating, even if the resulting doomsday machines have built-in uncertainties because of possible failures in C³I -- some of which may be irremediable -- and other factors (e.g., lack of will to order a second strike).

5. Conclusions

MAD is not only an acronym for "mutual assured destruction" but also for "mutual assured deterrence." In its former incarnation, MAD is more of an epithet than a statement of policy, except insofar as it implies that to save the world each side must be willing to destroy it.

Our Deterrence Equilibrium suggests this is only partially true: there is not, and need not be, "assured destruction," but only a probabilistic threat of such to induce "assured deterrence." If the threat of retaliation is sufficiently great, and perceived to be credible, neither side will find it advantageous to preempt.

Credibility depends on precommitments by both sides to implement a (probabilistic) threat, which certainly seems to characterize the nuclear retaliatory policies of the superpowers. As we indicated earlier, however, probabilistic threats of full-fledged retaliation may be interpreted as diminished responses to a provocation, but carried out with certainty.

Such responses in repeated play of a game would, it seems, drive one up the escalation ladder. Fortunately, the nuclear rung has never been reached in any superpower confrontation, which seems at least partially explained by the dynamic stability of the Deterrence Equilibrium -- after any perturbation in a player's preemption probability, that probability tends to be restored to zero. Thus equilibrium is maintained by a powerful force.

This self-restoring quality of the Deterrence Equilibrium will be reinforced by robust threats, which are always above the threshold level necessary to deter but never commit a player to retaliation for certain. Because these threats are both equally damaging and equally costly whatever one side thinks the other might do, they would, we believe, enhance the stability of the Deterrence Equilibrium in a game of incomplete information.

The difficult question to answer is what, operationally, constitutes a robust threat. We argued earlier that the present nuclear doctrines of the superpowers seem to preclude a certain response, except, perhaps, to a massive nuclear attack wherein all signs are unambiguous. On the other hand, they would seem to imply probabilities above the (minimal) threshold values. But are these threats, and the actions to make them credible, as nonprovocative as practicable?

If false signals should trigger an unprovoked attack, the consequences surely would be deadly. Since both superpowers have formidable second-strike capability from their largely invulnerable submarine-launched missiles, it seems better to err on the side of not being responsive enough -- having "only" a probabilistic threat, which our model indicates is quite sufficient -- rather than making one's retaliation too

automatic or too sensitive to provocation.

If it is hard to say exactly what constitutes a robust threat today, there is no ambiguity in our model about the undesirability of preemption. It is never optimal unless one can rest assured that the other side will never retaliate. Since this presumption seems hopelessly naive, there seems no good reason ever to contemplate preemption, given at least threshold threats of retaliation by both sides.

Yet this is not necessarily to commend "no first use" at levels below that of superpower confrontation. In response to a conventional attack, it is conceivable that holding out the possibility of introducing nuclear weapons into a conventional conflict may help to deter being attacked in the first place. But then this benefit must be weighed against the increased risk of nuclear escalation should the conflict actually occur and there be no self-imposed restraint on the first use of nuclear weapons.

This and other instances of potentially apocalyptic conflict that deterrence may prevent from erupting seem capable of game-theoretic modelling. At least in the case of the Deterrence Game, the effects of threats that underlie nuclear deterrence seem salutary. But when threats themselves become provocative and severely undermine trust, one must ask whether their deterrent value exceeds the costs of creating an inflammatory situation.

APPENDIX

In this appendix we shall conduct an exhaustive search for Nash equilibria in the Deterrence Game and analyze their properties. The rules of this game, along with payoff and strategy definitions, are given in section 3. The game is depicted in Figure 2.

The expected payoffs of Row (R) and Column (C) are repeated below:

$$E_R(s,q;t,p) = str_3 + (1-s)tp + s(1-t)qr_2; \quad (2)$$

$$E_C(t,p;s,q) = stc_3 + s(1-t)q + (1-s)tpc_2. \quad (3)$$

Our search will be broken down according to the values of s and t at the equilibrium.

Case 1: $s = t = 1$.

From (2), if $t = 1$ then $\frac{\partial E_R}{\partial s} = r_3 - p$. Since $s = 1$ at equilibrium only if $\frac{\partial E_R}{\partial s} \geq 0$, $p \leq r_3$ is necessary. Analogous consideration of (3) shows that $q \leq c_3$ at any equilibrium with $s = 1$. Now suppose that $t = 1$ and $p \leq r_3$. From (2), R's expected payoff is

$$E_R(s, q; 1, p) = p + s(r_3 - p),$$

so that R can never do better than to choose $s = 1$ and $q \leq c_3$. Similarly, $t = 1$, $p \leq r_3$ is C's best response to $s = 1$, $q \leq c_3$.

Therefore, the only equilibria consistent with Case 1 are

$$s = 1, q \leq c_3; t = 1, p \leq r_3. \quad (4)$$

The family (4) is called the Deterrence Equilibrium, since every strategy combination in the family leads to the same outcome -- the cooperative outcome of the underlying Chicken game, with payoffs (r_3, c_3) . Properties of the Deterrence Equilibrium will be adduced below.

Case 2: $t = 0$.

From (2), if $t = 0$ then

$$E_R(s, q; 0, p) = sqr_2,$$

so that R can maximize his expected value only by choosing $s = q = 1$.

If $s = q = 1$, then (3) shows that

$$E_C(t, p; 1, 1) = 1 - t(1 - c_3),$$

so that C's best choice is $t = 0$, and his payoff does not depend on p .

Therefore, the only equilibria consistent with Case 2 are

$$s = 1, q = 1; t = 0, p \text{ arbitrary,}$$

which we call the Preemption by C Equilibrium. At this equilibrium, the outcome of the Deterrence Game is always the outcome of the underlying Chicken game associated with a "win" for C -- the outcome with payoffs $(r_2, 1)$.

Case 3: $s = 0$.

This case is analogous to Case 2, and reduces to the Preemption by R Equilibrium:

$$s = 0, q \text{ arbitrary; } t = 1, p = 1.$$

The outcome corresponds, in Chicken, to a "win" for R, and has payoffs $(1, c_2)$.

Case 4: $0 < s < 1, t = 1$.

If $0 < s < 1$, (3) implies that $\frac{\partial E_C}{\partial p} = (1-s)tc_2 > 0$ provided $t > 0$.

Thus, at any equilibrium with $0 < s < 1$ and $t = 1, p = 1$ also since E_C is increasing in p . Now if $t = 1$ and $p = 1$,

$$E_R(s, q; 1, 1) = 1 - s(1-r_3)$$

by (2), so that R's expected payoff is maximized only when $s = 0$. This contradiction shows that there are no equilibria consistent with Case 4.

Case 5: $s = 1, 0 < t < 1$.

This case contains no equilibria, by an argument analogous to that for Case 4.

Case 6: $0 < s < 1$, $0 < t < 1$.

Equation (2) shows that $\frac{\partial E_R}{\partial q} = s(1-t)r_2$ so that, for an equilibrium with $0 < s < 1$ and $0 < t < 1$, $q = 1$ is a necessary condition since E_R is increasing in q . Analogously, so is $p = 1$. Now suppose that $0 < t < 1$ and $p = 1$ are fixed. To maximize

$$E_R(s, q; t, 1) = t + s[tr_3 - t + (1-t)qr_2],$$

it is clear that R must choose either $s = 0$ or $s > 0$ and $q = 1$. We discard $s = 0$ since it is not consistent with Case 6. Now in order that some s satisfying $0 < s < 1$ maximize E_R , it must be that $\frac{\partial E_R}{\partial s} = 0$, i.e.,

$$tr_3 - t + (1-t)r_2 = 0.$$

This equation implies that

$$t = t^* = \frac{r_2}{1 - r_3 + r_2}.$$

Note that $0 < t^* < 1$. Analogously, for fixed s and q satisfying $0 < s < 1$ and $q = 1$, $p = 1$ and some t satisfying $0 < t < 1$ maximize E_C only if

$$s = s^* = \frac{c_2}{1 - c_3 + c_2},$$

where, again, $0 < s^* < 1$. Finally, one can verify directly that

$$s = s^*, q = 1; t = t^*, p = 1 \tag{5}$$

is an equilibrium. We refer to this equilibrium as the Naive Equilibrium.

It is easy to show that, at the Naive Equilibrium, the players'

expected payoffs are

$$E_R^* = \frac{r_2}{1 - r_3 + r_2}, \quad E_C^* = \frac{c_2}{1 - c_3 + c_2},$$

and that $r_2 < E_R^* < r_3$ and $c_2 < E_C^* < c_3$. Thus, the Deterrence Equilibrium (4), with payoffs (r_3, c_3) , is Pareto-superior to the Naive Equilibrium (5).

The Deterrence Equilibrium possesses a dynamic-stability property which, once it forms, will (in repeated play) contribute to its persistence. To see this, assume that the Deterrence Equilibrium (4) has become established, and, further, that

$$0 < q < c_3, \quad 0 < p < r_3 \tag{6}$$

holds. Suppose that player C is concerned that there is some chance that R will preempt, i.e., that $s < 1$, and that C is therefore contemplating whether he should preempt with some positive probability. In other words, C is no longer sure that $s = 1$, and is reconsidering his choice of $t = 1$. But now differentiation of (3) yields

$$\frac{\partial E_C}{\partial t} = s(c_3 - q) + (1 - s)pc_2,$$

so that, if (6) holds, $\frac{\partial E_C}{\partial t} > 0$ for every value of s satisfying $0 \leq s \leq 1$. Therefore, C is motivated to choose $t = 1$, despite his doubts about the value of s , since E_C is increasing in t . A similar calculation shows that R is motivated to choose $s = 1$ regardless of his perception of the value of t , providing (6) holds. Thus, probabilistic threats of retaliation which are more than minimal ($q = c_3, p = r_3$) but less than certain ($p = 0, q = 0$) will tend to restore the Deterrence Equilibrium if it is perturbed.

FOOTNOTES

1. Steven J. Brams, Superpower Games: Applying Game Theory to Superpower Conflict (New Haven, CT: Yale University Press, forthcoming 1985), chaps. 1 and 2.
2. Steven J. Brams and Donald Wittman, "Nonmyopic Equilibria in 2 x 2 Games," Conflict Management and Peace Science 6, no. 1 (Fall 1981), pp. 39-62; D. Marc Kilgour, "Equilibria for Far-sighted Players," Theory and Decision 16, no. 2 (March 1984), pp. 135-157; see also Frank C. Zagare, "Limited-Move Equilibria in 2 x 2 Games," Theory and Decision 16, no. 1 (January 1984), pp. 1-19.
3. Steven J. Brams and Marek P. Hessel, "Threat Power in Sequential Games," International Studies Quarterly 28, no. 1 (March 1984), pp. 15-36.
4. Ibid.; the original distinction between compellent and deterrent threats is due to Thomas C. Schelling, Arms and Influence (New Haven, CT: Yale University Press, 1966).
5. For debate on this point, see Frank C. Zagare, "Toward a Reconciliation of Game Theory and the Theory of Mutual Deterrence" (Department of Political Science, Boston University, 1983); Brams and Hessel, "Threat Power in Sequential Games"; and Brams, Superpower Games, chap. 1.
6. Reinhard Selten, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," International Journal of Game Theory 4, no. 1 (1975), pp. 25-55; see also Martin Shubik, Game Theory in the Social Sciences: Concepts and Solutions (Cambridge, MA: MIT Press, 1982), pp. 265-270.
7. Thomas C. Schelling, The Strategy of Conflict (Cambridge, MA: Harvard University Press, 1960), chap. 9.
8. Brams, Superpower Games, chap. 1.

9. Paul Bracken, The Command and Control of Nuclear Weapons (New Haven, CT: Yale University Press, 1983). Gauthier claims that such precommitments are not necessary to deter aggression, but threats which are not credible are empty, and empty threats invite attack. His calculus of deterrence, we believe, is sensible only when his retaliator's threats will assuredly be implemented because of precommitments. See David Gauthier, "Deterrence, Maximization, and Rationality," Ethics 94, no. 3 (April 1984), pp. 474-495.

10. Brams, Superpower Games, pp. 45-46.