*Credibility, Uncertainty, and Deterrence**

D. Marc Kilgour, *Department of Mathematics, Wilfrid Laurier University*
Frank C. Zagare, *Department of Political Science, State University of New York at Buffalo*

In this paper the connection between deterrence stability and threat credibility is examined conceptually and theoretically. We formulate as a model of bilateral deterrence a game of incomplete information in which each player is uncertain about its opponent's preferences should it unilaterally alter the status quo. Uncertainty about the preferences of one's opponent leads to uncertainty about the opponent's willingness to retaliate. By identifying the credibility of each player's retaliatory threat with the probability that a player prefers retaliation to capitulation, we maintain consistency with both the traditional strategic literature, where credibility usually means believability, and with game theory, where credibility is usually synonymous with sequential rationality (i.e., subgame perfect equilibrium). We analyze formally the strategic implications of this conception of credibility and thus explore the critical role played by uncertainty in deterrence. By explicitly modeling uncertainty, we are able to understand the role of threats in contributing to, or detracting from, the robustness of a deterrence relationship.

Equilibrium was the name of the game.
—Henry Kissinger, *White House Years*

Credibility, Freedman (1981, 96) once observed, is the "magic ingredient" of deterrence. Never has this statement been more true than in the nuclear age when the capability of each superpower to inflict unacceptable damage on the other is evident.

In spite of its centrality to the theory of deterrence, the concept of credibility is seldom defined rigorously, let alone analyzed systematically, in the strategic literature. Most strategic analysts seem to be of the opinion that the term is transparent enough that no formal definition is required. Credibility of a threat to punish an aggressor is frequently taken to mean that the threat is believed, and left at that. Typically, the next analytical step is to explore the underlying determinants of such beliefs.

In the formal literature, there have been a few attempts to explore the connection between deterrence stability and threat credibility, but none of these at-

*D. Marc Kilgour gratefully acknowledges the financial support of the Natural Sciences and Engineering Research Council of Canada under grant A8974. Frank C. Zagare gratefully acknowledges the financial support of the United States Institute of Peace under grant USIP-532. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the United States Institute of Peace or the NSERC of Canada.

American Journal of Political Science, Vol. 35, No. 2, May 1991, Pp. 305–34

tempts has resulted in a thorough treatment. For example, Zagare (1987) links each player's credibility to his or her actual preferences but examines only those pure cases in which the player's threat is known to be (or not be) credible, thereby reducing credibility to a simple dichotomous variable. Brams and Kilgour (1988), in associating credibility with the probability that players believe a retaliatory threat will be executed, treat credibility continuously. But they model action choices only, so that the interplay of beliefs, uncertainty, and actions remains exogenous to their model.

In this essay we shall marry these approaches to analyze the role played by credibility in deterrence and crisis relationships, defining it in terms of the players' preferences and, at the same time, allowing each player to assess the credibility of an opponent's retaliatory threat probabilistically. This will permit us to analyze quantitatively the strategic implications of credibility and to explore the critical role played by uncertainty in deterrence. In our approach the principal source of uncertainty lies in lack of information about the preferences of one's opponent. By explicitly modeling uncertainty, we are better able to understand the role of threats in contributing to, or detracting from, the robustness of a deterrence relationship.

Alternate approaches to deterrence (e.g., Powell 1987; Nalebuff 1986) link it to the manipulation of the probability of an accidental war or to a lengthy learning sequence in which players try to demonstrate a greater capacity to absorb punishment (Powell 1988). In contrast, our view is that the essence of deterrence is the effort to discourage an opponent from choosing an explicitly preemptive action in a crisis by threatening retaliation against that action. Deterrence problems that are mutual are more complex: each side is trying to deter the other.

Credibility, Believability, and Rationality

In the strategic literature, credibility has usually been taken to be synonymous with believability (Freedman 1981; George and Smoke 1974; Jervis 1985; Schelling 1966); conversely, threats that are not believed are seen as incredible—as was the Eisenhower administration's threat to inflict nuclear devastation on the Soviet Union for relatively minor transgressions of the status quo. At the time the policy of Massive Retaliation was formulated, it was widely criticized for being unbelievable and, consequently, lacking credibility (Kaufmann 1956). As Smoke (1987, 88) put it, "The threat was not *credible* in the face of growing Soviet strategic power. As the Soviet arsenal of atomic bombs, and of long-range bombers to deliver them, grew during [the mid- to late 1950s], it became less believable that the United States would actually launch an atomic war over some invasion in Asia or elsewhere."

The credibility of threats has also been closely linked with their rationality.

Lebow (1981, 15), for one, notes that the difficulty of imparting credibility to the threat to go to war in the nuclear age stems from the fact that "the adversary knows the inherent *irrationality* of such threats" (emphasis added). Significantly, this connection between credibility and rationality is found not only in the strategic literature of deterrence but also in the game-theoretic literature, where the credibility of threats is generally taken to be synonymous with subgame perfectness of Nash equilibria, that is, with equilibria that are consistent with rational choices on all possible paths of the game tree (Selten 1975; Friedman 1986, 80–82; Holler 1988; Zagare 1990).

In the strategic literature, therefore, the notion of credibility is either directly or indirectly associated with rational or self-interested behavior. Credible threats are threats that are believed; threats can be believed exactly when they are rational to carry out; thus, only rational threats are credible. But what constitutes a rational threat? The answer to this vexing question depends on the way rationality is defined.[1] For our purposes there are two possible responses.[2]

One could identify "rational threats" by carefully delineating the real world conditions that would justify a retaliatory response by one nation to an untoward action of another, thereby separating those situations in which a deterrent threat is credible from those in which it is not. Gray (1979, 55), for example, does just this in arguing that a massive nuclear assault of the Soviet Union "would likely trigger a Soviet response in kind," since, under these circumstances, the Soviets would have "nothing left to lose." Curiously, while Gray seems to impute almost perfect credibility to the Soviet threat to respond to an all-out U.S. attack, he questions the credibility of the U.S. deterrent under similar circumstances, seeing little "merit (let alone moral justification) in executing the posthumous punishment of an adversary's society."

At the level of policy determination, speculation about the conditions under which an adversary would either contemplate an attack or respond to an initiative is certainly called for. But at the level of theory construction, such speculation is, for obvious reasons, counterproductive. Rather than enter into a debate about what would, and would not, precipitate an attack or a response by some nation—a question that in any case must ultimately be answered by those in policymaking positions—we take a second, more limited, approach to specifying credible threats. We simply define a credible threat to be one that the threatener would

[1]The lines of one debate about "rational" deterrence are sharply drawn in the January 1989 issue of *World Politics*. This symposium focuses on the relative usefulness of the case study approach and deductive methods for studying deterrence and is only indirectly related to the issues of rationality raised by this essay (see Achen and Snidal 1989; George and Smoke 1989; Jervis 1989; Lebow and Stein 1989; Downs 1989).

[2]For a discussion of the differences, implications, and compatibility of these two approaches to rationality, see Zagare (1990).

prefer to execute at the time it is to be executed.[3] Like Bueno de Mesquita (1981, 172) and Powell (1987, 719), we assume that an actor prefers to execute a threat when the expected worth of doing so exceeds the expected worth of failing to do so. Otherwise, the threat is irrational and, hence, incredible.[4]

Deterrence under Complete Information about Credibility

What role does credibility play in the strategic equation? How credible must a threat be in order to deter an opponent? What is the precise relationship between deterrence stability and the intrinsic issues of a crisis, the costs associated with outright conflict, and the value of the status quo? To answer these and related questions, we begin by exploring the three logically possible mutual deterrence games of complete information in which each player's threat is known to be either credible (i.e., rational) or not. In the next section, we develop a model in which the players are uncertain of the credibility of each other's threats. By analyzing that model, we are able to provide a coherent picture of the interrelationship of credibility, uncertainty, and deterrence.

Consider first the general outline of a typical deterrence problem in which each of two adversaries is trying to prevent the other from upsetting the status quo. To simplify matters, assume a two-stage game in which each side has two broad strategic choices: either to cooperate, *C*, with the other by supporting the status quo or not to cooperate, *D*, by attempting to overturn it. These actions give rise to four equally broad outcomes: if each state cooperates, the status quo (*CC*) reigns; if one side cooperates and the other does not, the latter gains an advantage (either *CD* or *DC*); and if neither side cooperates, conflict (*DD*) results (see Figure 1.A).

Mutual deterrence occurs when each side chooses *C* initially, presumably in the belief that, if it were to change its choice to *D* in the second stage, the opponent might well match that choice. Thus, a simple model of a crisis deterrence problem is an extensive game that ends after one simultaneous choice if

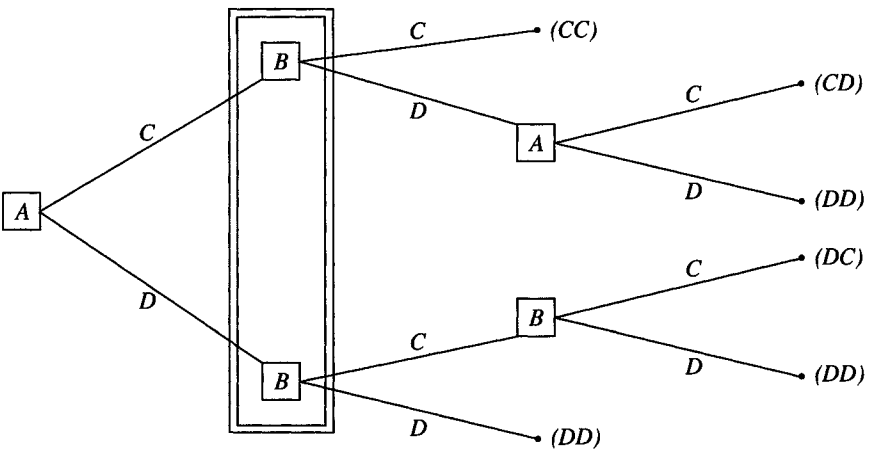
[3]Our model is of a single-play deterrence situation of which a nuclear crisis is perhaps the clearest example. In such a situation, it is unlikely that one player will attempt to build a reputation for firmness in order to influence the behavior of the other player in a similar game in the future because it is unlikely that there will be any (meaningful) future game with the same opponent once a nuclear sanction has been imposed. For the literature on reputation in repeated games, see, inter alia, Kreps and Wilson (1982); Wilson (1985); Sobel (1985); Wagner (1988); and Alt, Calvert, and Humes (1988).

[4]For the purpose of exploring the precise link between credibility and deterrence stability, we find this definition necessary. One could, of course, modify this definition of credible threats to permit the players to commit to a particular strategy at the start of the game. But as Rasmusen (1989, 94) points out, "Allowing commitment is the same as allowing equilibria to be non-perfect." For a discussion of the theoretical implications of precommitment strategies, see Brams and Kilgour (1988).

Figure 1.A. Possible Outcomes of a Deterrence Game

Nation A \ Nation B	Cooperate (*C*)	Not Cooperate (*D*)
Cooperate (*C*)	Status Quo (*CC*)	Advantage to *B* (*CD*)
Not Cooperate (*D*)	Advantage to *A* (*DC*)	Conflict (*DD*)

Figure 1.B. Complete Information Deterrence Game



Key:
□ = Decision node
• = Terminal node
C = Cooperate
D = Defect
▣ = Information partition

either *CC* or *DD* is attained initially, but offers the *C* player a second choice (to stay with *C* or to retaliate by changing to *D*) if *CD* or *DC* is reached in the first stage. This simple model is shown in Figure 1.B.

To develop Figure 1 into a model of mutual deterrence under complete information, it is necessary to specify only the players' preferences over the four possible outcomes. We always assume that each side prefers an advantage to the status quo and prefers the status quo to an advantage for the other side. This

assumption establishes a preference relationship among three of the four outcomes in the game. To wit,

for nation A: $DC >_A CC >_A CD$ and

for nation B: $CD >_B CC >_B DC$

where "$>_A$" means "is preferred by A to," and so on. We also assume that $CC >_A DD$, and $CC >_B DD$; this, too, is a reasonable assumption, since without it deterrence is not possible (Zagare 1987, chap. 4). All that remains to be specified,[5] therefore, is each player's preference between conflict DD and an advantage to the other side (either CD or DC).

For most deterrence games, these latter preference relationships are critical. The reason is simple: in this situation of complete information, they directly determine the credibility of the players' deterrent threat. For example, if a player prefers conflict to capitulation, then this preference is known, and therefore its threat to retaliate is credible (under our definition) because it is rational to carry out. But when this preference is reversed, the threat is irrational and therefore incredible by our definition.

With respect to this critical preference relationship, there are only three possibilities: each side prefers to execute the retaliatory threat; only one side does; or neither side does. In other words, both players, only one player, or no player possesses a credible retaliatory threat. These three situations are shown, respectively, in the two-by-two matrices depicted in Figure 2. Note that preferences are displayed as ordinal payoffs: each player's payoffs are denoted 1, 2, 3, 4 for its worst, next-worst, next-best, and best outcome respectively.

In Figure 2.A (Prisoner's Dilemma), each player prefers conflict to capitulation. Thus, each player's threat is *perfectly* credible as long as each has complete information about the other's preferences. Under these conditions mutual deterrence constitutes a stable relationship (Zagare 1987). (Of the two Nash equilibria in this game, mutual deterrence is strictly preferred by both players.) The reasoning behind this conclusion is straightforward: a player can anticipate that the opponent's retaliatory threat will be executed in response to a violation of the status quo. Since each player, by assumption, prefers the status quo to conflict, each is deterred. Consequently, deterrence is stable when it is rational for each player to carry out a retaliatory threat.

What are the strategic implications of an asymmetry of credibility, implying the preferences shown in Figure 2.B ("Called Bluff")?[6] Using reasoning similar

[5] We also ignore complexities due to power asymmetries and other variations of each player's evaluation of the status quo. For an examination of the strategic consequences of these considerations, see Zagare (1987).

[6] In Figure 2.B nation B's threat is credible and nation A's is not. The analysis to follow applies both to this situation and to its mirror image, in which the players' roles are reversed.

Figure 2. Three Sets of Preferences over Deterrence Game Outcomes

Nation A	Nation B: Cooperate (C)	Nation B: Not Cooperate (D)
Cooperate (C)	Status Quo (3, 3)	Advantage to B (1, 4)
Not Cooperate (D)	Advantage to A (4, 1)	Conflict (2, 2)

A. "Prisoner's Dilemma"

Nation A	Nation B: Cooperate (C)	Nation B: Not Cooperate (D)
Cooperate (C)	Status Quo (3, 3)	Advantage to B (2, 4)
Not Cooperate (D)	Advantage to A (4, 1)	Conflict (1, 2)

B. "Called Bluff"

Nation A	Nation B: Cooperate (C)	Nation B: Not Cooperate (D)
Cooperate (C)	Status Quo (3, 3)	Advantage to B (2, 4)
Not Cooperate (D)	Advantage to A (4, 2)	Conflict (1, 1)

C. "Chicken"

to the above, it is easy to show that when only one player possesses a credible threat, the outcome associated with that player's best outcome is the unique Nash equilibrium. Hence, deterrence is not stable, and the player with the credible threat will gain the advantage. The player whose opponent lacks a credible threat will move rationally to upset the status quo by choosing *D* because the opponent, preferring capitulation to conflict, will rationally accede to this action. Thus, the player with the credible retaliatory threat "wins."

Deterrence is also unstable when neither player's retaliatory threat is credible, as illustrated by Figure 2.C ("Chicken"). Each player prefers to be the only one to upset the status quo, since the other, lacking a credible retaliatory threat, would not respond rationally. The outcomes associated with an advantage for *A*

or for B are therefore the two perfect Nash equilibria of the game. In other words, when both players prefer capitulation to conflict and when threats must be credible in the sense of being rational (as defined above), deterrence is unstable, and the outcome is uncertain. Mutual deterrence is not an equilibrium.

The real world, of course, is not so simple or transparent. It is characterized by, among other things, nuance, ambiguity, equivocation, duplicity, and ultimately uncertainty. Typically, policymakers are unable to acquire complete information about the intentions of their opponents; at best, they can hope to obtain probabilistic knowledge of these key determinants of interstate behavior. Gray's (1979) speculation about the "likely" Soviet response to various American initiatives (see above) is an example. Clearly, Gray's analytic uncertainty stems from lack of information about Soviet preferences. As just demonstrated, with rational actors and complete information, the success of mutual deterrence in a crisis situation is easy to determine.

The same insight applies equally to other long-standing strategic conundrums. Would the United States risk Washington or New York for Paris or Bonn, as de Gaulle once asked rhetorically? Given the uncertain nature of interstate politics, the only possible answer to this question is perhaps. What about Toronto? Maybe. San Francisco or Atlanta? Probably. But not certainly. If one could answer certainly (or certainly not) to such questions, a deterrence game would be completely specified. Given complete information, players would know with which of the situations of Figure 2 they are faced and could act accordingly. (Actions may not be completely determined by this information, but the stability of mutual deterrence certainly is.) When players, however, have only uncertain knowledge of an opponent's preferences, they cannot tell whether the game they are playing is like "Chicken," "Called Bluff," or Prisoner's Dilemma. Of course, a player knows his or her own preferences and can eliminate some possibilities, but uncertainty remains integral to most real world deterrence situations, a fact that is strategically crucial.

Yet players with real life deterrence problems can, and do, make estimates about the likely motives (and actions) of an opponent. Sometimes these estimates are correct, and sometimes, like Chamberlain's about Hitler's intentions in 1938, or like Hitler's about British intentions in 1939, they are not. Such guesses concern the likely preferences of the opponent and, consequently, the probable structure of the game in which the players are involved.

Behind these estimates lies what Joynt and Corbett (1978, 94–95) call the "curve of credibility." This curve "begins with defense of the homeland, descends to clearly defined spheres of influence or the territory of allies and then drops to near zero for the defense of other interests." In other words, since a state will be more likely to respond to certain kinds of incursions than to others, its credibility will vary across the range of conflict issues.

What is significant about the "curve" of credibility, however, is not its

existence but its shape. Since credibility obviously has important implications for deterrence in an uncertain world, we next develop a formal model that delineates its strategic implications. In the end we aim to specify explicitly the connection between mutual deterrence, preferences, and threat credibility.

Deterrence and Uncertainty

To model the role played by uncertainty in bilateral strategic relationships, we postulate the three-stage extensive game of incomplete information shown by Figure 3. As before there are two players, A and B, with the same two choices and, consequently, the same four outcomes as given previously (see Figure 4). The payoffs are now in (von Neumann-Morgenstern cardinal) utilities: A receives 1 at DC, a_3 at CC, and 0 at CD; and B receives 1 at CD, b_3 at CC, and 0 at DC. We also assume that both players are aware of these payoffs that are always consistent with the ordinal ranking of any mutual deterrence game—that is, each player prefers his or her own advantage to the status quo, and the status quo to an advantage for the other. But unlike the games previously examined, our model postulates that each player's relative preferences for conflict, DD, are known only probabilistically to the opponent. The players' utilities at DD are A_2 (to A) and B_2 (to B), where A_2 and B_2 are independent binary random variables with known distributions. Specifically, it is common knowledge that

$$A_2 = \begin{cases} a_{2^+} \text{ with probability } p_A \\ a_{2^-} \text{ with probability } 1 - p_A \end{cases}$$

$$B_2 = \begin{cases} b_{2^+} \text{ with probability } p_B \\ b_{2^-} \text{ with probability } 1 - p_B \end{cases}$$

so that both players know these distributions, both know that they know, and so on. However, only A knows the actual value of A_2 and only B knows the actual value of B_2; the realization of these values constitutes the first stage of the game. All parameter values are fixed and satisfy

$$0 \leqslant p_A \leqslant 1$$

$$0 \leqslant p_B \leqslant 1$$

$$a_{2^-} < 0 < a_{2^+} < a_3 < 1$$

$$b_{2^-} < 0 < b_{2^+} < b_3 < 1$$

The credibility of each player's threat to retaliate will depend on the particular value of p_A and p_B. When this value is high, the threat will be seen to be more credible than not credible. But when this value is low, it will be less credible. Overall, A's (respectively, B's) threat will be credible and its preferences like those in Prisoner's Dilemma with probability p_A (p_B); A's (B's) threat will

Figure 3. Incomplete Information Deterrence Game

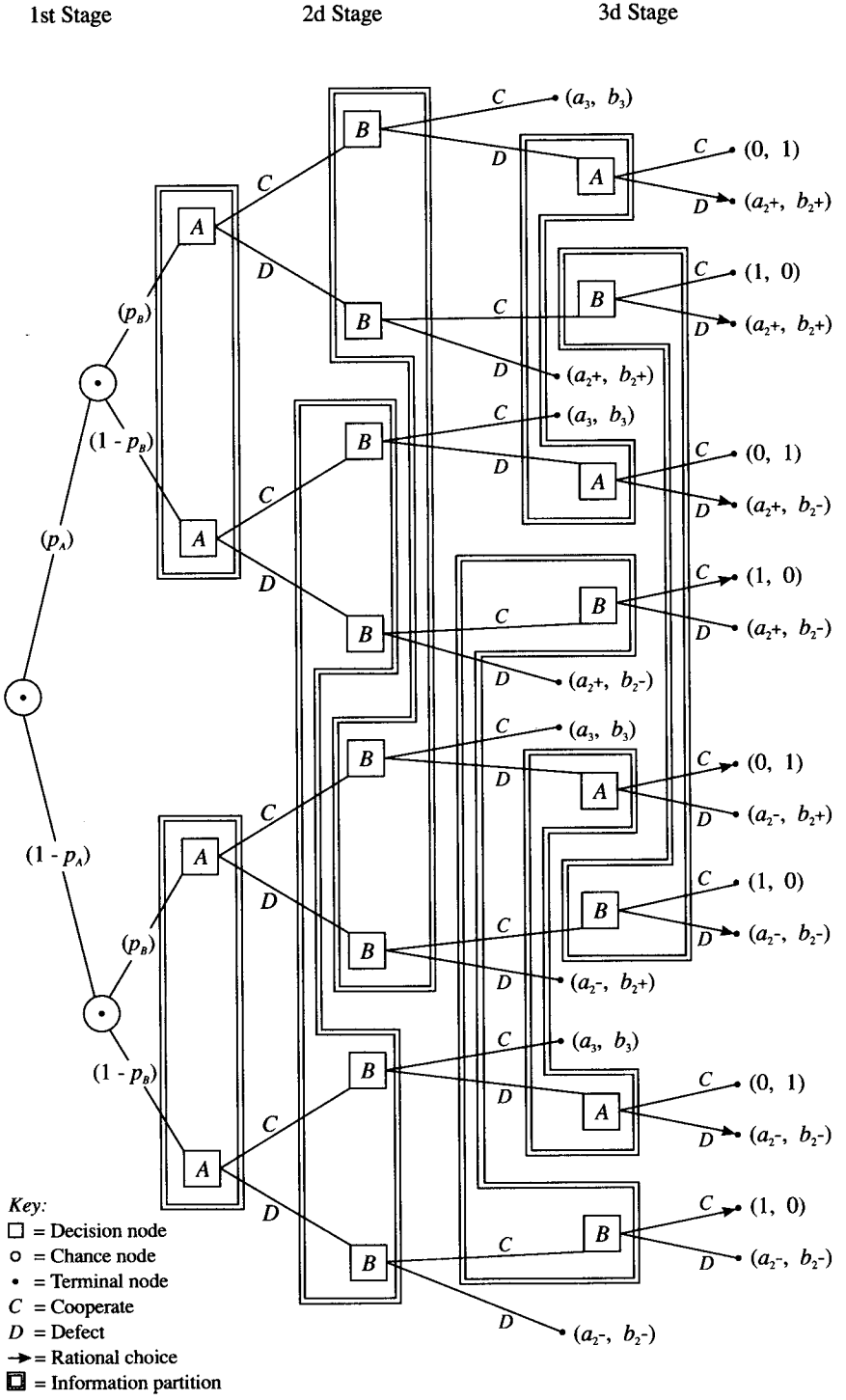


Figure 4. Preferences over Incomplete Information Deterrence Game Outcomes

		Nation B: Cooperate (C)	Nation B: Not Cooperate (D)
Nation A	Cooperate (C)	Status Quo (a_3, b_3)	Advantage to B $(0, 1)$
	Not Cooperate (D)	Advantage to A $(1, 0)$	Conflict (A_2, B_2)

be incredible and its preferences like those in "Chicken" with probability $1 - p_A$ $(1 - p_B)$.

Following the determination of the *DD* payoffs that define threat credibility in the first stage, the players simultaneously[7] choose either *C* or *D* in stage 2. They next learn which cell (*CC, CD, DC,* or *DD*) their stage 2 choices imply. A player who has cooperated while the opponent defected will then get an opportunity to change his or her selection, that is, to retaliate. For example, if the outcome of the second stage is *CD, A* may then either stay at that outcome (remain with *C*) or retaliate (change to *D*), thereby moving the game from *CD* to *DD*. Similarly, should the stage 2 outcome be *DC, B* will have the option of staying there or moving to *DD*. The game ends after the simultaneous selection of strategies and, if applicable, a subsequent retaliation decision by either *A* or *B*.

The process we attempt to model, then, goes like this: first, something exogenous happens (first stage) which determines, a priori, the probability that each player will prefer to respond to a noncooperative act by the other. We remain silent, however, on exactly what this "something" might be—perhaps a significant change in the underlying power relationship of the two nations (Organski and Kugler 1980); or an internal power shift like the one that occurred in the Soviet military prior to the 1962 Cuban Missile Crisis (Allison 1971); or a shock to the environment like the 1973 Middle East war that embroiled the two superpowers in a tricky deterrence game. Thus, this model applies equally well to immediate or to general deterrence games or even to international crises.

In any event, after the (first) stage has been set, each player makes a strategic choice based upon its own evaluation of the outcomes (i.e., the stakes) and its estimate of the preferences of its opponent. This choice constitutes the second

[7]As preliminary to a full extension of this model to bilateral deterrence relationships, we assume that the players make simultaneous moves at the start of the game. For situations in which there is no obvious first mover, such as the Middle East crisis between the superpowers in 1973 (see Zagare 1987 for a discussion), this assumption is warranted. However, it may sometimes be the case that one player is clearly in the position of defending the status quo, and the other of challenging it. In such situations the assumption of sequential choice is more appropriate. For a discussion of the differences and strategic consequences implied by these two assumptions, see Kilgour and Zagare (1990).

stage of our model. For example, in June 1948 Soviet decision makers decided to clamp a blockade around Berlin in the belief that the Western powers would have little choice but to accept this as a fait accompli. The Soviets were obviously mistaken, however, since in the third stage of this particular deterrence game the Western powers chose to resist the Soviet move by launching an airlift of food and other supplies to Berlin. Such a retaliation decision, if applicable, constitutes the third and final stage of the game.

For the purpose of clarifying the payoffs of this game, assume for the moment that each player *believes* he or she has complete information about an opponent's preferences, that is, that p_A and p_B equals either 1 or 0. (In other words, ignore for now the information sets drawn around the various nodes of Figure 3.) Then only one route through the first stage has positive probability, and one of the deterrence games of complete information (see Figures 1.B and 2) must arise. For example, it may now happen that each nation *believes* that the other prefers to retaliate if the status quo is violated. Based upon this belief, mutual deterrence is a very likely possibility, since it is the unique Pareto-optimal equilibrium; the payoffs to nation A and nation B will then be a_3 and b_3, respectively. But this particular outcome, *CC*, can actually be achieved in four distinct ways. Reading the (a_3, b_3) final outcomes in Figure 3 from the top, each player may have been correct in his or her estimate, A may have been incorrect and B correct, B incorrect and A correct, or both incorrect in assessing the rationality of the other's threat.

Next, suppose that B estimates (incorrectly) that A prefers to "chicken out" if B upsets the status quo and that A (also incorrectly) estimates that B possesses a credible threat. The payoffs to the players at *DD* would then be (a_{2^+}, b_{2^-}). This outcome derives from the following sequence of decisions: A, believing B's threat to be credible, is deterred and chooses C. Simultaneously, B, believing A's threat to be incredible, chooses D and induces, temporarily, outcome *CD*. Then in the third stage, A, consistent with its actual preferences, retaliates, thereby inducing a payoff of a_{2^+} for itself and b_{2^-} for B. This is the sequence leading to the upper (a_{2^+}, b_{2^-}) outcome in Figure 3.

Finally, consider one of the cases discussed in the previous section in which neither player has a credible threat, and both know this. This is the last (lowest) final outcome in Figure 3. Under these conditions of accurate information *and* simultaneous choice, mutual deterrence is surely unstable. The payoffs to A and B would be a_{2^-} and b_{2^-}, respectively.

The outcomes represented by the other endpoints of the tree can be similarly interpreted. Of course, the particular endpoint reached and the consequent payoffs to the players depend on each player's preferences and perception, correct or not, of his or her opponent's preferences.

Parenthetically, it is interesting to point out that so far our simple model has revealed that deterrence can succeed when both sides correctly gauge each other's intentions, or even when one or both players misperceive the environment. In

addition, deterrence failure may result from either a correct or an incorrect reading of the strategic situation. We therefore can make the following observations:

> Misperception is neither necessary nor sufficient for the failure of mutual deterrence.
>
> Accurate assessments of the strategic environment are neither necessary nor sufficient for the success of mutual deterrence.[8]

Analysis

But what happens when the players do not know for sure what the opponent will do in the case of a defection?[9] In other words, what are the strategic implications of incomplete information about the credibility of the other's threat? How credible must a threat be to deter? What is the precise connection between the credibility and magnitude of the retaliatory threat? To answer these questions, we analyze in general the incomplete information game of Figure 3, considering in particular the information sets surrounding some of the nodes of the tree. Hence, we take the parameters p_A and p_B to have values somewhere between 0 and 1.

We begin our analysis by using backward induction on the last stage of the game. This stage is easy to analyze because a decision maker always has complete information over his or her own payoffs. For example, A may know that the game is at point CD and that $A_2 = a_{2^+}$; A's uncertainty over whether $B_2 = b_{2^+}$ or $B_2 = b_{2^-}$ is of no strategic significance to the decision to retaliate (resulting in DD) or to stay (resulting in CD). A's choice is based solely on A's own payoffs; in this example, A receives a_{2^+} at DD versus only 0 at CD and therefore chooses retaliation, causing the game to end at DD. By similar reasoning all the third-stage choices can be determined and are indicated by arrows on the tree of Figure 3.

Because third-stage behavior can be determined, the players' strategic decisions are reduced to the (second-stage) choice of C or D, which *does* depend on the state of knowledge of the player making the selection. Following the conventions of game theory regarding mixed strategies, let

x_{PD} = probability nation A chooses C given that $A_2 = a_{2^+}$

x_{Ch} = probability nation A chooses C given that $A_2 = a_{2^-}$

y_{PD} = probability nation B chooses C given that $B_2 = b_{2^+}$

y_{Ch} = probability nation B chooses C given that $B_2 = b_{2^-}$

[8] Bueno de Mesquita and Lalman (1988) make similar observations.

[9] See Harsanyi (1967–68) for the seminal treatment of games of incomplete information. A useful summary is found in the Appendix of Tirole (1988).

These probabilities can be thought of as each player's "cooperation" policy for each of the two situations that can arise in a deterrence situation: either the player's preferences are like those in Prisoner's Dilemma—in which case the player will prefer to retaliate if challenged (i.e., $A_2 = a_{2^+}$ or $B_2 = b_{2^+}$) and will cooperate initially with probability x_{PD} or y_{PD}—or its preferences are like those in "Chicken" (i.e., $A_2 = a_{2^-}$ or $B_2 = b_{2^-}$) so that it will prefer to capitulate rather than retaliate and will cooperate initially with probability x_{Ch} or y_{Ch}. Thus, the two types of nation A choose strategies x_{PD} and x_{Ch}, and the two types of nation B choose strategies y_{PD} and y_{Ch}.

The expected payoffs to each type of each player can now be determined. If A is of type PD, then A's expected payoff is $E_{A|PD} = E_{A|PD}(x_{PD}; y_{PD}, y_{Ch})$, where

$$\begin{aligned} E_{A|PD}(x_{PD}; y_{PD}, y_{Ch}) = {} & p_B[x_{PD}y_{PD}a_3 + (1 - x_{PD}y_{PD})a_{2^+}] \\ & + (1 - p_B)[x_{PD}y_{Ch}a_3 + (1 - x_{PD})y_{Ch} \\ & + (1 - y_{Ch})a_{2^+}] \end{aligned}$$

If A is of type Ch, A's expected payoff is

$$\begin{aligned} E_{A|Ch}(x_{Ch}; y_{PD}, y_{Ch}) = {} & p_B[x_{Ch}y_{PD}a_3 + (1 - x_{Ch})a_{2^-}] \\ & + (1 - p_B)[x_{Ch}y_{Ch}a_3 + (1 - x_{Ch})y_{Ch} \\ & + (1 - x_{Ch})(1 - y_{Ch})a_{2^-}] \end{aligned}$$

Nation B's expected payoffs, $E_{B|PD}(y_{PD}; x_{PD}, x_{Ch})$ and $E_{B|Ch}(y_{Ch}; x_{PD}, x_{Ch})$ are analogous.

Next, we inquire which of the strategy combinations $(x_{PD}, x_{Ch}; y_{PD}, y_{Ch})$ are in equilibrium. Because in the third stage any player who must act has a strictly dominant choice, the Bayesian equilibria of the game with payoffs $E_{A|PD}$, $E_{A|Ch}$, $E_{B|PD}$, and $E_{B|Ch}$ are sequential equilibria as well as perfect Bayesian equilibria (see Tirole 1988). In particular, this means that, at each equilibrium, each player always acts optimally in accordance with his or her beliefs and revises those beliefs rationally, in accordance with the actions of the opponent. These are the natural equilibrium concepts for our study because they highlight the interrelationship of credibility and actions.[10]

The equilibria of the incomplete information deterrence game are determined in the Appendix. It turns out that all but 14 strategy combinations can be eliminated immediately. Of these 14, four are transitional equilibria that occur under only a very limited set of conditions and shall not be discussed here.[11] This

[10]The equilibria obtained are also trembling-hand perfect, that is, limits of equilibria that apply if there is a small probability that both players will choose every possible strategy (Selten 1975; Tirole 1988).

[11]Existence conditions for the four transitional equilibria, which we label $E2$, $E3$, $E12$, and $E13$, are given in the Appendix.

leaves but 10 perfect Bayesian equilibria, or stable strategy combinations, to be described. (For a listing, see Table A.1 in the Appendix.)

For our purposes the most interesting of the 10 equilibria is the one we call the *sure-thing deterrence equilibrium.* Our immediate task will be to describe the conditions necessary for its existence; after this, we shall group it and the remaining equilibria into four categories in order to interpret the results of our formal analysis.

The Sure-Thing Deterrence Equilibrium

The sure-thing deterrence equilibrium (or $E1$ in the Appendix) is the strategy combination [1, 1; 1, 1], that is, $x_{PD} = x_{Ch} = y_{PD} = y_{Ch} = 1$. This equilibrium implies the choice of C by each player regardless of whether he or she prefers retaliation or capitulation. And since each player's policy is never to attack (preempt) the other, no matter what its retaliation/nonretaliation preferences may be, peace is at hand.

Obviously, the sure-thing deterrence equilibrium is robust; when it exists, the status quo is secure. Unfortunately, it occurs under conditions that may be more restrictive than any of the other equilibria we have identified. Specifically, it is demonstrated in the Appendix that, for sure-thing deterrence to be a perfect Bayesian equilibrium, both of the following inequalities must hold:

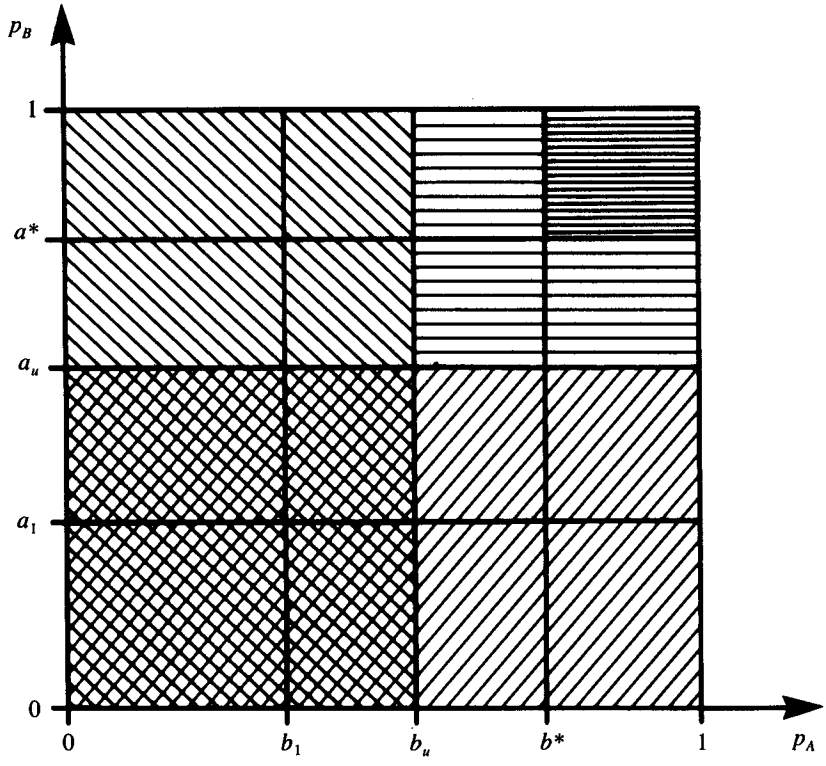
$$p_B \geq a^* = \frac{1 - a_3}{1 - a_{2^+}} \qquad p_A \geq b^* = \frac{1 - b_3}{1 - b_{2^+}}$$

To understand the strategic significance of these inequalities, it is best to refer to Figure 5. The horizontal axis of this figure represents p_A, or the probability that $A_2 = a_{2^+}$; similarly, the vertical axis represents p_B, the probability that nation B's preferences are like those of Prisoner's Dilemma, that is, that $B_2 = b_{2^+}$. Along these two axes are indicated several constants, such as a^* and b^*, that are defined and discussed in the Appendix. These constants are convenient thresholds for categorizing and interpreting the equilibria of the incomplete information deterrence game.

Notice from Figure 5 that both a^* and b^* are plotted fairly close to $p_A = 1$ and $p_B = 1$, respectively. Thus, the sure-thing deterrence equilibrium is to be found only in the extreme northeast region of the figure, above and to the right of the threshold values. That these two parameters are close to 1 means that for this particular equilibrium to occur, each side must place a relatively large probability on the other's willingness to retaliate against an attempt to upset the status quo, that is, each side's credibility must be high. But perfect credibility, $p_A = p_B = 1$, is *not* necessary for deterrence stability.

Going beyond the obvious, notice that a^* and b^* are defined in terms of the

Figure 5. Incomplete Information Deterrence Game: Location of Equilibria



Key:

Class 1 Equilibria

Equilibrium $E1$ (Class 1)

Class 2_A Equilibria

Class 2_B Equilibria

Classes 2_A 2_B, and 3 Equilibria

magnitude of each player's evaluation of the status quo (a_3 and b_3) *and* of a_{2^+} and b_{2^+}, which are the values of the outcome associated with mutual conflict when the players prefer conflict to capitulation. We observe that

> As the values of a_3 and b_3 increase, the threshold values a^* and b^* decrease. This means that, ceteris paribus, the higher each player's evaluation of the existing distribution of values (i.e., the status quo), the more likely the sure-thing deterrence equilibrium exists.

This relationship between deterrence and satisfaction with the status quo can be seen as a theoretical justification for Henry Kissinger's policy of detente. By drawing the Soviet Union into a more lucrative economic and political relationship, the United States could, in principle, reduce the credibility requirements necessary for deterring noncooperative Soviet policies. Or, stated differently, when the value of the status quo is increased, deterrence is more likely.

> As the values of a_{2^+} and b_{2^+} decrease, the threshold values a^* and b^* also decrease. This means that, ceteris paribus, deterrence stability is also enhanced by increasing the costs associated with mutual punishment.

Here we find clearly the principal device used to bolster a policy of deterrence. By increasing the costs of warfare, deterrence becomes more likely as the credibility requirements of a deterrent threat become less onerous. In other words, if the costs of mutual punishment go up so that its value goes down, then deterrence can be achieved with less credible threats. This is fortunate indeed, since there is almost certainly an inverse relationship between the credibility of threats and the costs associated with their execution.

Significantly, however, our analysis reveals that there is a point at which such a strategem becomes ineffective and even counterproductive. Note that the sure-thing deterrence equilibrium depends only on the value of a_{2^+} and b_{2^+} and not on a_{2^-} and b_{2^-}. Thus, given that p_A and p_B are fixed, if $A_2 = a_{2^-}$ or $B_2 = b_{2^-}$, then deterrence is not made more probable by further decreases in the utility (payoff) of the conflict outcome.[12] Moreover, if p_A and p_B are not fixed, then mutual deterrence may be more difficult to establish if increases in the cost of warfare actually decrease the perceived value of p_A and p_B. In the extreme, when p_A and p_B are near zero so that neither side possesses a credible threat, there is no equilibrium that involves either individual or mutual restraint. And as we describe shortly, mutual deterrence is not likely in between, when $p_A < b^*$ and $p_B < a^*$.

[12]In the late 1950s, both the United States and the Soviet Union lacked a secure second-strike capability. Under these conditions our deduction about an inverse relationship between the costs of warfare and deterrence stability remain intact. When neither side has a deliverable deterrent threat, the credibility of threats is spurious and deterrence problematic because of "the reciprocal fear of surprise attack" (Schelling 1960).

Thus, we find little theoretical support for the argument of Intriligator and Brito (1984) and other classical deterrence theorists that mutual overkill adds to the stability of deterrence. Rather, our model suggests that for crises in core areas where credibility is naturally high, a minimum deterrent strategy that relies "on the retention of only enough nuclear weapons to provide an assured destruction capability" (Kegley and Wittkopf 1989, 351) is better. Such a policy would not make a retaliatory threat incredible but would impose significant costs on an aggressor and therefore would suffice for deterrence stability.

> As the value of a_{2^+} approaches a_3, and as b_{2^+} approaches b_3, a^* and b^* approach 1. This means that, ceteris paribus, deterrence is less likely as each player's evaluation of the outcome associated with mutual punishment gets closer to the value of the status quo. Under these conditions, a player will be almost as satisfied fighting as not fighting and therefore will be more likely to defect from cooperation.

The connection between these two parameter values (a_{2^+} and a_3, or b_{2^+} and b_3) can perhaps shed some theoretical light on the reasoning behind the Japanese attack on Pearl Harbor in 1941. Historians seem to agree that this particular breakdown in the international system did not occur because Japan placed a high probability on a military victory; rather, Japanese leaders believed that the United States might not fully resist, and even if it did and subsequent events were unfavorable to Japan, suffering the consequences of a military defeat was not much worse than enduring an unsatisfactory and humiliating status quo. Parenthetically, we might add that it is not necessary to assume, as Snyder and Diesing (1977, 124–27) do, that Japan actually preferred war to the status quo in order to explain its behavior. In an uncertain world, deterrence might collapse, and a war might occur when the players' relative evaluations of these two outcomes are close.

The relative closeness of these two values, at least for U.S. decision makers, might also explain the buildup in both conventional and nuclear forces in the early years of the Reagan administration. One explanation offered for this buildup is that it was driven not so much by the structural determinants that lie behind many arms races (Baugh 1984) as by the perception of the U.S. president that the weakness of the Soviet economy would eventually cause the Soviets to falter and perhaps even drop out altogether (Bailer and Afferica 1982–83). In other words, as the value of the mutual punishment outcome (i.e., a defeated and demoralized Soviet Union) approached the value of the existing security regime for the United States, the less likely it became that the United States would cooperate. Contrariwise, one might further speculate, as did many in the 1988 presidential campaign, that it was the cost to the Soviets imposed by this buildup that induced later Soviet cooperation in negotiating the Intermediate Nuclear Force (INF) treaty. This assertion depends upon the assumption that the

Soviets were committed to keeping pace with any and all Western strategic initiatives. Our model shows that this proviso is necessary, for if they preferred otherwise, then the imposition of additional costs would be unrelated to the calculus of deterrence.

Finally, we should note that the sure-thing deterrence equilibrium might just as well have been labeled the "cooperate regardless" equilibrium, since it involves initial and unconditional cooperation by both players at the start of the game. Since its existence depends upon each player's perception that the other will retaliate with a very high probability should either player not cooperate initially, this equilibrium, and the strategies associated with it, can be considered a one-shot game analogue of the tit-for-tat equilibrium strategies identified by Axelrod (1984). (For a theoretically congruent connection between Axelrod's findings and tit-for-tat strategies in single play games, see Zagare 1987, 27.)

Other Equilibria

As indicated, the sure-thing deterrence equilibrium is but one of 10 perfect Bayesian equilibria that occur with positive probability. These 10 equilibria can be grouped into four distinct categories. We next present these categories and discuss the implications for deterrence of the existence conditions summarized in Table A.1 in the Appendix.

Class 1: Good-chance equilibria. There are three nontransitional equilibria in class 1, which consists of equilibria of the form $[\cdot, 1; \cdot, 1]$, where $(\cdot)$ signifies any value. In an equilibrium of this class, therefore, a player may preempt the opponent if he or she in fact prefers to retaliate if challenged; that is, if his or her preferences are like those in Prisoner's Dilemma. If not, then the player's preferences are like those in "Chicken" (i.e., $A_2 = a_{2^-}$ or $B_2 = b_{2^-}$), and he or she will cooperate with certainty.

Clearly, the sure-thing deterrence equilibrium is a member of this class. The other two equilibria, which we here and in the Appendix call $E4$ and $E5$, are real and interesting possibilities in our uncertain deterrence game.

$E4$ consists of the strategy combination $[u, 1; v, 1]$, which means that if nation A (respectively, B) finds that its preferences are like those of Prisoner's Dilemma, then it will attack nation B (A) with probability $1 - u(1 - v)$; $E5$ is the combination $[0, 1; 0, 1]$, which can be interpreted similarly. The only difference between $E4$ and $E5$ is that in $E5$ each player will preempt the other with certainty (rather than probabilistically) should it find that it would prefer to retaliate if preempted itself. In both cases a player who prefers capitulation to confrontation will always cooperate.

Summarizing all three cases, then, a player with "Chicken"-like preferences will be deterred if he or she perceives that the probability that the other will retaliate is high enough. Conversely, if at least one player is willing to endure the costs of mutual punishment, deterrence can, but need not, fail.

As discussed in the Appendix, the existence conditions for $E5$ are somewhat different from those for the sure-thing deterrence equilibrium ($E1$) and $E4$, so that it is possible for $E5$ to be the unique equilibrium. (Figure 5 shows this case.) If this happens, deterrence will emerge only if *both* players prefer not to retaliate if attacked. At the same time, however, each player must convince the other that there is a sufficiently high probability that he or she in fact prefers retaliation to capitulation. Under the circumstances, this would be no mean feat. Consequently, deterrence is more problematic when the only equilibrium is $E5$ than when $E1$ and $E4$ also exist.

On the other hand, if the sure-thing deterrence equilibrium ($E1$) exists, then $E4$ also exists, and $E5$ generally does too. Fortunately, when this occurs, the sure-thing deterrence equilibrium is Pareto superior (in fact, it is strictly preferred by both types of both players) to either of the other two (see Appendix). Thus, as long as sure-thing deterrence is an equilibrium, there are compelling theoretical reasons to expect that a robust deterrence regime will emerge and endure.

That said, it is tempting to speculate about which of these equilibria have been characteristic of the actual strategic relationship of the superpowers, and at which times. But this would be unduly speculative. Even though the strategies associated with each of these three deterrence equilibria are different, the implied action choices are not necessarily distinguishable. For example, the festering crisis over Berlin during the latter days of the Eisenhower administration may not have erupted into an all-out superpower conflict because each nation selected a strategy consistent with sure-thing deterrence, or because each selected a mixed strategy consistent with $E4$, or because $E5$ was in play, and each state was able to convince the other that, contrary to its true preferences, it was willing to resist any attempt by the other to alter the status quo (Zagare 1987, 165–66). On the other hand, our model does reveal that for deterrence to have worked in Berlin in 1958 and elsewhere, whatever the actual preferences of the two powers, each must have attained more than a modicum of threat credibility, for otherwise, deterrence is unstable and peace unlikely.

Classes 2_A and 2_B: No-chance equilibria. These two sets of three equilibria are mirror images of each other. Class 2_A consists of equilibria of the form [0, 0; ·, 1], while class 2_B corresponds to [·, 1; 0, 0]. In class 2_A, which is composed of $E6$, $E7$, and $E8$, nation A's strategy is always to defect; this is similarly true of nation B in class 2_B, which consists of the equilibria $E9$, $E10$, and $E11$ (see Appendix). Since all six of these equilibria involve certain defection by one player (the "preemptor") and certain submission by the other when it (the "nonpreemptor") prefers not to retaliate, there is no chance of deterrence succeeding under any of these equilibria.

The equilibria within each class differ, however, with respect to the non-

preemptor's *initial* policy given that he or she prefers retaliation to submission, that is, when it is the Prisoner's Dilemma type. As indicated in Table A.1, in this event, B's strategy associated with $E6$ and A's with $E9$ involve the choice of C with certainty; both $E7$ and $E10$ involve the probabilistic choice of C by the nonpreemptor; and $E8$ and $E11$ are associated with the certain choice of D. No matter what the second player's initial choice, however, the status quo will be violated. Furthermore, conflict may occur unless the nonpreemptor prefers to capitulate.

As one might expect, equilibria of this kind exist when *either* p_A *or* p_B is sufficiently low. This means that a calculated rupture of deterrence becomes more likely as the probability increases that at least one player prefers to "chicken out"; this provides the other player with an incentive to risk conflict in order to gain an advantage. Specifically, for equilibria of class 2_A or 2_B to exist, either $p_A \leqslant b_u$ or $p_B \leqslant a_u$. As indicated in the Appendix, these threshold values depend on the values of b_3 and b_{2^-} and a_3 and a_{2^-}, respectively. Thus, for small negative values of either b_{2^-} and a_{2^-}, the region of no-chance deterrence equilibria covers almost the entire square of Figure 5. Under these conditions, the breakdown of deterrence is very likely. Conversely, as the values of b_{2^-} or a_{2^-} decline (i.e., the costs associated with this unwanted payoff increase), the zone of existence of class 2 equilibria shrinks.

In other words, in conflicts where at least one player's retaliatory threat is not very credible, increasing the cost of unwanted conflict makes deterrence more likely. Under such conditions, therefore, nuclear wespons may have a salutary effect: as the price of guessing wrong increases, a state is less likely, ceteris paribus, to upset the status quo.

Class 3: The slim-chance equilibrium. But what if *both* p_A and p_B are low so that neither player's threat is particularly credible? Under these conditions, an additional equilibrium outcome, here called the "slim-chance" equilibrium but in the Appendix referred to as $E14$, may be possible in addition to the equilibria of classes 2_A and 2_B. (The slim-chance equilibrium and two transitional equilibria, which share the form [0, ·; 0, ·], compose class 3.)

The slim-chance equilibrium takes the form [0, u; 0, v]. This means that the choice of a player who prefers conflict over capitulation is to attack immediately. On the other hand, when a player prefers capitulation over conflict, he or she chooses to preempt with a certain specific probability. (See Appendix for details.)

There is, perhaps, one positive feature of the slim-chance equilibrium. Notice from Figure 5 (or Table A.1) that if this equilibrium exists there is a fairly high probability that each player actually prefers capitulation to confrontation. Because there is a reasonably high probability that the players have "Chicken"-like preferences, there is also certainly some positive probability that they will

choose *not* to preempt. Thus, deterrence can emerge when each player's credibility is low, but this emergence may be as much a matter of luck as strategic structure.

Finally, the slim-chance equilibrium always coexists with both classes of no-chance equilibria (classes 2_A and 2_B). As we indicate in the Appendix, all are Pareto-optimal, so that there is no natural way to single out one equilibrium as most likely to be selected by rational agents. Hence, when both players lack a particularly credible threat, there is a slim chance of deterrence; it is much more likely, however, that deterrence will fail—the outcome in this case is uncertain.

Summary and Conclusions

In this paper we examined the theoretical connections between deterrence stability and threat credibility. We have done so by formulating as a model of mutual deterrence a game of incomplete information in which each player is uncertain about how his or her opponent prefers to respond should the player unilaterally alter the status quo. By identifying the credibility of each player's threat to retaliate with the probability that a player prefers retaliation to capitulation, we maintain consistency with both the traditional strategic literature, in which credibility is usually equated with believability, and with the literature of game theory, in which credibility is usually taken to be synonymous with rationality.

Perhaps the signal contribution of our model is to provide a measure of the circumstances under which deterrence can emerge in an uncertain world. Specifically, when the credibility of each player's threat is sufficiently high, deterrence is very likely, though perhaps not certain, as some deterrence theorists have speculated (e.g., Intriligator and Brito 1981, 256; Mueller 1989). Credibility thresholds for the existence of the three "good-chance" deterrence equilibria are dependent on each player's evaluation of the status quo and the costs associated with mutual conflict. Contrary to Lebow (1984, 181) and other deterrence theorists, however, we found no linear or other simple relationship between the costs of warfare and deterrence stability. In fact, our model indicates that in core areas, where both players have inherently credible threats, increasing the costs of mutual punishment past a certain point does little to enhance deterrence stability. And if there is, as we suspect, an inverse relationship between these costs and threat credibility, then increasing the costs of war at this level makes deterrence less likely, not more likely.

For this reason we recommend for the security of each superpower's homeland policies of deterrence that are sufficient to inflict unacceptable damage on an opponent yet are survivable enough to be available for a *retaliatory* attack. At the strategic level, therefore, we favor arms reductions, single-warhead missiles, and hardening of silos. A thin "defensive" system around second-strike forces is also consistent with the spirit of our findings.

By contrast, in conflict areas where the credibility of *both* players is lower,

our model suggests a different policy. As the players' curves of credibility drop in tandem, deterrence stability *is* enhanced as the costs of deterrence failure are increased. In Europe, for instance, the inherently lower credibility of each superpower's "extended" deterrent threat can potentially be offset by increasing the costs of guessing wrong. From the perspective of our model, therefore, the recently negotiated removal of intermediate nuclear forces from Europe by the superpowers has reduced the robustness of the prevailing equilibrium in that area of the world. On the other hand, the existence of independent French and British nuclear forces operating in an arena of high inherent credibility probably makes such a reduction immaterial.

Much the same can be said of conflict in peripheral areas where an asymmetry of credibility exists. Under these conditions, deterrence of an unsatisfied player is most unlikely. Such instability can be mitigated, to some extent, by increasing the damage that retaliation can wreak. But, in the limit, if one player's credibility becomes negligible, then deterrence cannot survive. Recent superpower initiatives in Afghanistan and Grenada are testimony to this harsh reality.

It is interesting to observe that the history of the postwar period roughly conforms to the predictions of our model. In core areas where credibility is high, deterrence has indeed prevailed, and war has been avoided. The principal breakdowns of deterrence have come in areas in which one superpower or the other has a vested interest and consequently a higher level of credibility. When deterrence failed in Hungary or Czechoslovakia because of Soviet actions, or in Vietnam because of a U.S. decision to reestablish a deteriorating status quo, the locale was such that the core interests of the offended superpower were not at risk.

To be sure there are exceptions to this statement. But it is telling that the exceptions include all of the dramatic cases in which strategic deterrence almost evaporated. In Cuba, for instance, the Soviet Union directly challenged the interests of a stronger United States and, not surprisingly, eventually backed down. In Berlin, starting in 1948 and continuing until the mid-1960s, persistent Soviet challenges very nearly upset the foundations of the European equilibrium. Finally, the two superpowers came close to war again during the 1973 Middle East War when the Soviets threatened to intervene in order to protect Egypt's Third Army and President Sadat's pro-Soviet regime. It is consistent with the deductions of our model to hypothesize that in the latter two cases, where each side was more or less equally motivated, deterrence prevailed because the threats of relatively severe retaliation offset the concomitant decline of each side's reduced credibility.

If strategic deterrence has been the rule, and small breakdowns of deterrence the exception, since 1945, the question arises of how the strategies pursued by the superpowers brought this state of affairs into being. Or, put another way, what is the nature of the equilibrium that has characterized superpower behavior since the dawning of the nuclear age? It is possible, both empirically and logically, for each superpower sometimes to have selected strategies consistent with

the sure-thing deterrence equilibrium: never behave aggressively, but threaten with high credibility a harsh retaliatory strike. But at other times, actual deterrence and observable behavior are also consistent with other equilibrium strategies that entail the possibility of a probe and a deterrence failure. The most disturbing of these is clearly the "slim-chance" equilibrium where stability is more a function of luck than anything else. Obviously, our formal analysis can shed no light on transcendental questions such as which deterrence equilibrium is now in play, or how equilibria can be changed; but it is interesting to note that many strategic thinkers who are skeptical about the persistence of mutual deterrence (e.g., Jones and Thompson 1978) consider good fortune to be a necessary part of the dynamic of deterrence.

Manuscript submitted 12 September 1989
Final manuscript received 1 May 1990

APPENDIX

This appendix contains the detailed analysis of the game defined in the text (see Figure 3). Recall that player A's strategic variables are x_{PD} and x_{Ch} and that player B's are y_{PD} and y_{Ch}. The eight parameter values satisfy $0 \leq p_A \leq 1$, $0 \leq p_B \leq 1$, $a_{2^-} < 0 < a_{2^+} < a_3 < 1$, and $b_{2^-} < 0 < b_{2^+} < b_3 < 1$. A's expected payoffs are

$$\begin{aligned} E_{A|PD}(x_{PD}; y_{PD}, y_{Ch}) &= p_B[x_{PD}y_{PD}a_3 + (1 - x_{PD}y_{PD})a_{2^+}] \\ &\quad + (1 - p_B)[x_{PD}y_{Ch}a_3 + (1 - x_{PD})y_{Ch} + (1 - y_{Ch})a_{2^+}], \qquad (1) \\ E_{A|Ch}(x_{Ch}; y_{PD}, y_{Ch}) &= p_B[x_{Ch}y_{PD}a_3 + (1 - x_{Ch})a_{2^-}] + (1 - p_B)[x_{Ch}y_{Ch}a_3 \\ &\quad + (1 - x_{Ch})y_{Ch} + (1 - x_{Ch})(1 - y_{Ch})a_{2^-}] \end{aligned}$$

and B's expected payoffs are analogous.

Taking the parameter values as fixed, we will find all Nash equilibria of this game. Our search is greatly aided by the following four lemmas.

LEMMA 1.

(a) $\dfrac{\partial E_{A|PD}}{\partial x_{PD}} \geq 0$ iff $X_{PD} = p_B(a_3 - a_{2^+})y_{PD} - (1 - p_B)(1 - a_3)y_{Ch} \geq 0$,
with equality iff $X_{PD} = 0$;

(b) $\dfrac{\partial E_{A|Ch}}{\partial x_{Ch}} \geq 0$ iff $X_{Ch} = p_B(a_3 y_{PD} - a_{2^-}) - (1 - p_B)[(1 - a_3 - a_{2^-})y_{Ch} + a_{2^-}] \geq 0$,
with equality iff $X_{Ch} = 0$.

PROOF: Simply differentiate relation (1) and combine terms. QED

To illustrate the usefulness of Lemma 1, suppose that a specific strategy combination $[x_{PD}, x_{Ch}; y_{PD}, y_{Ch}]$ is an equilibrium. Notice that the value of X_{PD} depends only on B's strategic variables, not on A's. If it happens that $X_{PD} > 0$, then $x_{PD} = 1$ is necessary at equilibrium. Also $X_{PD} < 0$ implies $x_{PD} = 0$, whereas $X_{PD} = 0$ is consistent with any value of x_{PD}. Similar inferences can be drawn about the relations of X_{Ch} with x_{Ch}, Y_{PD} with y_{PD}, and Y_{Ch} with y_{Ch}, where Y_{PD} and Y_{Ch} are defined analogously.

Some direct consequences of Lemma 1 give important information about the structure of equilibria.

LEMMA 2: Suppose that $[x_{PD}, x_{Ch}; y_{PD}, y_{Ch}]$ is an equilibrium.

If $x_{PD} > 0$, then $x_{Ch} = 1$. If $x_{Ch} < 1$, then $x_{PD} = 0$.

If $y_{PD} > 0$, then $y_{Ch} = 1$. If $y_{Ch} < 1$, then $y_{PD} = 0$.

PROOF: All of the conclusions will follow from Lemma 1 if it can be shown that

$$X_{Ch} > X_{PD} \quad \text{and} \quad Y_{Ch} > Y_{PD}. \tag{2}$$

But $X_{Ch} - X_{PD} = p_B y_{PD}\, a_{2+} - (1 - p_B)(1 - y_{Ch})\, a_{2-} > 0$ and analogously for $Y_{Ch} - Y_{PD}$. QED

LEMMA 3: Suppose that $[x_{PD}, x_{Ch}; y_{PD}, y_{Ch}]$ is an equilibrium. If $y_{PD} = 0$ and $y_{Ch} > 0$, then $x_{PD} = 0$. If $x_{PD} = 0$ and $x_{Ch} > 0$, then $y_{PD} = 0$.

PROOF: If $y_{PD} = 0$ and $y_{Ch} > 0$, then $X_{PD} < 0$, so $x_{PD} = 0$ follows from Lemma 1. The proof of the second assertion is analogous. QED

LEMMA 4: Suppose that $[x_{PD}, y_{Ch}; y_{PD}, y_{Ch}]$ is an equilibrium. If $y_{Ch} = 0$, then $x_{Ch} = 1$. If $x_{Ch} = 0$, then $y_{Ch} = 1$.

PROOF: If $y_{Ch} = 0$, $X_{Ch} = p_B a_3 y_{PD} - a_{2-} > 0$, since $a_{2-} < 0$. Therefore, $x_{Ch} = 1$ follows from Lemma 1. The second assertion is proven analogously. QED

It is a consequence of Lemmas 2, 3, and 4 that the only possible equilibria are the 14 combinations shown in Table A.1. (Note: In Table A.1, and throughout this Appendix, u denotes a value of a strategic variable of A satisfying $0 < u < 1$; v denotes a value of a strategic variable of B satisfying $0 < v < 1$.) Other symbols appearing in Table 1 are defined in the text, or below.

In terms of the game model described in the text, the interpretation of Lemma 2 is simple: at

Table A.1. Equilibria and Existence Conditions

		Strategic Variables				Existence Conditions		
Class	Equilibrium	x_{PD}	x_{Ch}	y_{PD}	y_{Ch}	on p_A		on p_B
1	$E1$	1	1	1	1	$p_A \geq b^*$	and	$p_B \geq a^*$
	$E2$ (T)	u	1	1	1	$p_A \geq b^*$	and	$p_B = a^*$
	$E3$ (T)	1	1	v	1	$p_A = b^*$	and	$p_B \geq a^*$
	$E4$	u	1	v	1	$p_A \geq b^*$	and	$p_B \geq a^*$
	$E5$	0	1	0	1	$p_A \geq b_u$	and	$p_B \geq a_u$
2_A	$E6$	0	0	1	1			$p_B \leq a_1$
	$E7$	0	0	v	1			$p_B < a_u$
	$E8$	0	0	0	1			$p_B \leq a_u$
2_B	$E9$	1	1	0	0	$p_A \leq b_1$		
	$E10$	u	1	0	0	$p_A < b_u$		
	$E11$	0	1	0	0	$p_A \leq b_u$		
3	$E12$ (T)	0	u	0	1			$p_B = a_u$
	$E13$ (T)	0	1	0	v	$p_A = b_u$		
	$E14$	0	u	0	v	$p_A \leq b_u$	and	$p_B \leq a_u$

Note: (T) = Transitional.

equilibrium a player is at least as aggressive when he or she is of the Prisoner's Dilemma type as when of the "Chicken" type. This is hardly surprising, because a PD-type player is not so averse to the conflict outcome that aggressiveness might bring about. Lemma 3 indicates that if your opponent is always aggressive with Prisoner's Dilemma–type payoffs but not always with "Chicken"-type payoffs, then you should always be aggressive if you are of the Prisoner's Dilemma type (because you may be able to take advantage of "chickening out"). Lemma 4 implies that a player whose opponent is always aggressive in "Chicken" should never be aggressive when he is the "Chicken" type—a conclusion that is easy to accept on the basis of the complete information game.

To simplify expressions for the existence for the 14 possible equilibria in this game, define

$$a_1 = \frac{1 - a_3}{1 - a_{2^-}}, \qquad a_u = \frac{1 - a_3}{1 - a_3 - a_{2^-}}, \qquad a^* = \frac{1 - a_3}{1 - a_{2^+}}$$

$$b_1 = \frac{1 - b_3}{1 - b_{2^-}}, \qquad b_u = \frac{1 - b_3}{1 - b_3 - b_{2^-}}, \qquad b^* = \frac{1 - b_3}{1 - b_{2^+}}$$

Note that $0 < a_1 < a_u < 1$, $a_1 < a^* < 1$, and $a^* \geq a_u$ iff $a_{2^+} - a_{2^-} \geq a_3$, with equality iff $a_{2^+} - a_{2^-} = a_3$; and similarily for b_1, b_u and b^*.

We now begin a systematic process which will determine existence conditions for each possible equilibrium. First suppose that $y_{PD} = y_{Ch} = 1$. It is easy to check that $X_{PD} \geq 0$ iff $p_B \geq a^*$. By relation (2), $p_B \geq a^*$ implies $X_{Ch} > 0$, and now Lemma 1 shows that A's best response is $x_{PD} = 1$ and $x_{Ch} = 1$. Carrying out an analogous calculation for B completes the justification of $E1$ of Table A.1: the combination [1, 1; 1, 1] is an equilibrium iff $p_B \geq a^*$ and $p_A \geq b^*$.

Again assuming that $y_{PD} = y_{Ch} = 1$, it is easy to verify that $X_{Ch} \leq 0$ if $p_B \leq a_1$. By relation (2), $X_{Ch} \leq 0$ implies $X_{PD} < 0$, and Lemma 1 now shows that A's best response is $x_{PD} = 0$ and $x_{Ch} = 0$. Now assume that $x_{PD} = x_{Ch} = 0$ and consider B's best response. It is easy to verify that $Y_{PD} = 0$ and $Y_{Ch} = -b_{2^-} > 0$, so Lemma 1 shows that B can do no better than to choose $y_{PD} = 1$ and $y_{Ch} = 1$. Thus [0, 0; 1, 1] is an equilibrium iff $p_B \leq a_1$, as indicated for $E6$ of Table A.1; the justification for $E9$ is analogous.

The calculations supporting the other assertions about pure strategy equilibria in Table 1 are similar. In particular,

$E5$: [0, 1; 0, 1] is an equilibrium iff $p_B \geq a_u$ and $p_A \geq b_u$

$E8$: [0, 0; 0, 1] is an equilibrium iff $p_B \leq a_u$

$E11$: [0, 1; 0, 0] is an equilibrium iff $p_A \leq b_u$

Because of Lemma 2, the game has no pure strategy equilibria other than $E1$, $E5$, $E6$, $E8$, $E9$, and $E11$.

To begin the analysis of equilibria involving exactly one mixed strategy, assume again that $y_{PD} = y_{Ch} = 1$. It is easy to verify that $X_{PD} = 0$ iff $p_B = a^*$. Therefore, by relation (2) and Lemma 1, A's best response can be $x_{PD} = u$ (where $0 < u < 1$) and $x_{Ch} = 1$, only if $p_B = a^*$. Now if $x_{PD} = u$ and $x_{Ch} = 1$, $Y_{PD} = p_A(b_3 - b_{2^+})u - (1 - p_A)(1 - b_3)$ so that $Y_{PD} \geq 0$ iff

$$u \geq \frac{(1 - p_A)(1 - b_3)}{p_A(b_3 - b_{2^+})} = u^* \tag{3}$$

and, of course, $Y_{Ch} > 0$ if $Y_{PD} \geq 0$, by relation (2). Furthermore, $u^* \leq 1$ iff $p_A \geq b^*$. Thus, necessary conditions for $E2$ of Table A.1, [u, 1; 1, 1] are $p_B = a^*$, $p_A \geq b^*$, and relation (3). It can be verified that these conditions are also sufficient. The situation for $E3$ is analogous.

Now suppose that $y_{PD} = y_{Ch} = 0$. Because $X_{PD} = 0$ and $X_{Ch} > 0$, $x_{PD} = u$ (where $0 < u < 1$) and $x_{Ch} = 1$ is a best response for A. Now assume that $x_{PD} = u$, $x_{Ch} = 1$. By relation (2), $y_{PD} = 0$,

and $y_{Ch} = 0$ will be a best response for B iff $Y_{Ch} \leq 0$. But $Y_{Ch} = p_A(b_3u - b_{2^-}) - (1 - p_A)(1 - b_3) \leq 0$ iff

$$u \leq \frac{(1 - p_B)(1 - a_3) + p_B a_{2^-}}{p_B a_3} = u_u \tag{4}$$

Furthermore, $u_u > 0$ iff $p_B < a_u$. This shows that $E10$ of Table A.1, $[u, 1; 0, 0]$, is an equilibrium iff both $p_B < a_u$ and relation (4) holds. The analysis for $E7$ is parallel.

If $y_{PD} = 0$ and $y_{Ch} = 1$, then $X_{PD} < 0$, and $X_{Ch} = 0$ iff $p_B = a_u$. By Lemma 1, $x_{PD} = 0$ and $x_{Ch} = u$ (where $0 < u < 1$) constitute A's best response. Now if $x_{PD} = 0$ and $x_{Ch} = u$, $Y_{PD} < 0$ and $Y_{Ch} = -p_A b_{2^-} - (1 - p_A)[(1 - b_3 - b_{2^-})u + b_{2^-}] \geq 0$ iff

$$u \leq \frac{-b_{2^-}}{(1 - p_A)(1 - b_3 - b_{2^-})} = u_l \tag{5}$$

Since $u_l > 0$ by assumption, $E12$ of Table A.1, $[0, u; 0, 1]$, is an equilibrium iff both $p_B = a_u$ and relation (5) holds. $E13$ is similar, completing the analysis of equilibria involving exactly one mixed strategy.

To analyze $E4$ of Table A.1, assume that $y_{PD} = v$ and $y_{Ch} = 1$, where $0 < v < 1$. In order that $x_{PD} = u$ and $x_{PD} = 1$ be A's best response, it is necessary and sufficient that $X_{PD} = 0$ by relation (2). But $X_{PD} = p_B(a_3 - a_{2^+})v - (1 - p_B)(1 - a_3) = 0$ iff

$$v = \frac{(1 - p_B)(1 - a_3)}{p_B(a_3 - a_{2^+})} = v^*$$

(Compare with relation 3.) Now $v^* > 0$ by assumption, and $v^* < 1$ iff $p_B > a^*$. By symmetry, it follows that equilibrium $E4$, $[u, 1; v, 1]$, occurs iff $p_A > b^*$, $p_B > a^*$, $u = u^*$, and $v = v^*$.

A similar argument can be used to prove that $E14$, $[0, u; 0, v]$, occurs iff $p_A < b_u$, $p_B < a_u$, $u = u_t$, and $v = v_t$, where $v_t = -a_{2^-}/[(1 - p_B)(1 - a_3 - a_{2^-})]$. (Compare relation 5.) This completes the justification of the necessary and sufficient conditions for existence given in Table A.1.

It is evident from inspection of Table A.1 that there is generally a multiplicity of equilibria corresponding to each particular set of parameter values. Fortunately some of these equilibria can be interpreted as unlikely to occur because others are preferred, often strictly preferred, by both players, and these preferences are independent of type. We now carry out these payoff comparisons for concurrent equilibria.

If $p_A \geq b^*$ and $p_B \geq a^*$, both $E1$ and $E4$ exist. It is easy to verify that the players' payoffs at $E1$ are $E_{A|PD}(E1) = a_3$, $E_{A|Ch}(E1) = a_3$, $E_{B|PD}(E1) = b_3$, and $E_{B|Ch}(E1) = b_3$.

To calculate A's payoffs at $E4$, note that the first equation of (1) can be rewritten in the form $E_{A|PD} = X_{PD}x_{PD} + W$. Because $x_{PD} = u^*$ at $E4$, it follows from Lemma 1 that $X_{PD} = 0$. This fact combined with $y_{PD} = v^*$ and $y_{Ch} = 1$ yields

$$E_{A|PD}(E4) = p_B a_{2^+} + (1 - p_B) \tag{6}$$

It follows easily that $E_{A|PD}(E1) - E_{A|PD}(E4) = (1 - a_{2^+})(p_B - a^*) \geq 0$, with equality iff $p_B = a^*$. Analogously, the second equation of (1) can be used to obtain $E_{A|Ch}(E4) = p_B v^* a_3 + (1 - p_B)a_3$, which implies that $E_{A|Ch}(E1) - E_{A|Ch}(E4) = p_B a_3(1 - v^*) \geq 0$, with equality iff $v^* = 1$, which occurs iff $p_B = a^*$. In summary, both types of A prefer $E1$ to $E4$ whenever they coexist (i.e., whenever $p_B \geq a^*$), and both types strictly prefer $E1$ to $E4$ when $p_B > a^*$. Analogous inferences can be drawn for B, leading to the conclusion that both types of both players strictly prefer $E1$ to $E4$ whenever $p_B > a^*$ and $p_A > b^*$; if $p_B > a^*$ and $p_A = b^*$, or $p_B = a^*$ and $p_A > b^*$, one player prefers $E1$ to $E4$, and the other is indifferent; and if $p_B = a^*$ and $p_A = b^*$, both players are indifferent.

As is clear from Table A.1, $E1$ and $E4$ are not the only equilibria when p_A and p_B are large. These two equilibria can coincide with $E5$, [0, 1; 0, 1]. It is easy to verify from equation (1) that $E_{A|PD}(E5) = p_B a_{2^+} + (1 - p_B)$, so (refer to relation 6) $E_{A|PD}(E1) \geq E_{A|PD}(E5)$ when $p_B \geq a^*$, and $E_{A|PD}(E1) = E_{A|PD}(E5)$ when $p_B = a^*$. Similarly $E_{A|Ch}(E5) = p_B a_{2^-} + (1 - p_B)a_3$, which implies that $E_{A|Ch}(E1) > E_{A|Ch}(E5)$ for all $p_B \geq a^*$. Analogous results can be obtained for B, confirming that $E1$ is Pareto superior to $E5$ when these equilibria coexist. In particular, when both p_A and p_B are close to 1, the unique Pareto-optimal equilibrium is $E1$.

When p_B is small, equilibria of the form [0, 0; y_{PD}, 1] arise—specifically $E6$, $E7$, and $E8$. (See Class 2_A in Table A.1.) It is easy to check that neither player's expected payoff ever depends on y_{PD} in this circumstance. (In fact, the path through the extensive game tree, Figure 3, does not actually depend on y_{PD}.) Therefore, both types of both players are indifferent among all equilibria of the form [0, 0; y_{PD}, 1]. The situation is similar for all equilibria of Class 2_B, specifically $E9$, $E10$, and $E11$, which exist when p_A is small.

When $p_A \leq b_u$ and $p_B \leq a_u$, three groups of equilibria coexist: Class 2_A; Class 2_B; and $E14$ in Class 3. With lengthy calculation it can be shown that

$$E_{A|PD}(2_B) < E_{A|PD}(E14) < E_{A|PD}(2_A)$$

$$E_{A|Ch}(2_B) < E_{A|Ch}(E14) < E_{A|Ch}(2_A)$$

and, analogously,

$$E_{B|PD}(2_A) < E_{B|PD}(E14) < E_{B|PD}(2_B)$$

$$E_{B|Ch}(2_A) < E_{B|Ch}(E14) < E_{B|Ch}(2_B)$$

Thus, the players' preferences over these equilibria are always exactly opposite, and all three are always Pareto optimal.

Finally, it is possible for $E1$ to coexist with equilibria of Class 2_A when $a_l < a^* \leq a_u$ and with equilibria of Class 2_B when $b_l < b^* \leq b_u$. It is easy to verify that both types of A always prefer $E1$ to equilibria of Class 2_B when they coexist and that if A is of type Ch, then A prefers $E1$ to equilibria of Class 2_A. If A is of type PD, A prefers $E1$ to equilibria of Class 2_A if $p_B > a^*$ and has the reverse preference if $p_B < a^*$.

REFERENCES

Achen, Christopher H., and Duncan Snidal. 1989. "Rational Deterrence Theory and Comparative Case Studies." *World Politics* 41:143–69.

Allison, Graham T. 1971. *Essence of Decision: Explaining the Cuban Missile Crisis*. Boston: Little, Brown.

Alt, James E., Randall L. Calvert, and Brian Humes. 1988. "Reputation and Hegemonic Stability: A Game-Theoretic Analysis." *American Political Science Review* 82:445–66.

Altfeld, Michael F. 1985. "Uncertainty as a Deterrence Strategy: A Critical Assessment." *Comparative Strategy* 5:1–26.

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Bailer, Seweryn, and Joan Afferica. 1982-83. "Russia and Reagan." *Foreign Affairs* 61:249–71.

Baugh, William H. 1984. *The Politics of Nuclear Balance*. New York: Longman.

Brams, Steven J. 1983. *Superior Beings: If They Exist, How Would We Know?* New York: Springer-Verlag.

Brams, Steven J., and D. Marc Kilgour. 1988. *Game Theory and National Security*. New York: Basil Blackwell.

Bueno de Mesquita, Bruce. 1981. *The War Trap*. New Haven: Yale University Press.

Bueno de Mesquita, Bruce, and David Lalman. 1988. "Arms Races and the Opportunity for Peace." *Synthese* 76:263–83.

Cioffi-Revilla, Claudio. 1983. "A Probability Model of Credibility." *Journal of Conflict Resolution* 27:73–108.

Downs, George W. 1989. "The Rational Deterrence Debate." *World Politics* 41:225–37.

Freedman, Lawrence. 1981. *The Evolution of Nuclear Strategy.* New York: St. Martin's Press.

Friedman, James W. 1986. *Game Theory with Applications to Economics.* New York: Oxford University Press.

George, Alexander L., and Richard Smoke. 1974. *Deterrence in American Foreign Policy.* New York: Columbia University Press.

———. 1989. "Deterrence and Foreign Policy." *World Politics* 41:170–82.

Gray, Colin S. 1979. "Nuclear Strategy: The Case for a Theory of Victory." *International Security* 4:54–87.

Harsanyi, John C. 1967–68. "Games with Incomplete Information Played by 'Bayesian' Players." 3 pts. *Management Science* 14 (series A): 159–82, 320–34, 486–502.

Holler, Manfred J. 1988. "Three Characteristic Functions and Tentative Remarks on Credible Threats." Memo 1988-1, Institute of Economics, University of Aarhus.

Intriligator, Michael D., and Dagobert L. Brito. 1981. "Nuclear Proliferation and the Probability of Nuclear War." *Public Choice* 37:247–60.

———. 1984. "Can Arms Races Lead to the Outbreak of War?" *Journal of Conflict Resolution* 28:63–84.

Jervis, Robert. 1985. "Introduction: Approach and Assumptions." In *Psychology and Deterrence,* ed. Robert Jervis, Richard Ned Lebow, and Janice Gross Stein. Baltimore: Johns Hopkins University Press.

———. 1989. "Rational Deterrence: Theory and Evidence." *World Politics* 41:183–207.

Jones, T. K., and W. Scott Thompson. 1978. "Central War and Civil Defense." *Orbis* 22:681–712.

Joynt, Carey B., and Percy E. Corbett. 1978. *Theory and Reality in World Politics.* Pittsburgh: University of Pittsburgh Press.

Kaufmann, William. 1956. "The Requirements of Deterrence." In *Military Policy and National Security,* ed. William Kaufmann. Princeton: Princeton University Press.

Kilgour, D. Marc, and Frank C. Zagare. 1990. "Asymmetric Deterrence." Presented at the annual meeting of the International Studies Association, Washington, DC.

Kegley, Charles W., and Eugene Wittkopf. 1989. *The Nuclear Reader: Strategy, Weapons, War.* 2d ed. New York: St. Martin's Press.

Kreps, David M., and Robert Wilson. 1982. "Reputation and Imperfect Information." *Journal of Economic Theory* 27:253–79.

Lebow, Richard Ned. 1981. *Between Peace and War: The Nature of International Crisis.* Baltimore: Johns Hopkins University Press.

———. 1984. "Windows of Opportunity: Do States Jump through Them?" *International Security* 9:147–86.

Lebow, Richard Ned, and Janice Gross Stein. 1989. "Rational Deterrence Theory: I Think, Therefore, I Deter." *World Politics* 41:208–24.

Mueller, John. 1989. *Retreat from Doomsday: The Obsolescence of Major War.* New York: Basic Books.

Nalebuff, Barry. 1986. "Brinkmanship and Nuclear Deterrence: The Neutrality of Escalation." *Conflict Management and Peace Science* 9:19–30.

Organski, A. F. K., and Jacek Kugler. 1980. *The War Ledger.* Chicago: University of Chicago Press.

Powell, Robert. 1987. "Crisis Bargaining, Escalation, and MAD." *American Political Science Review* 81:717–35.

———. 1988. "Nuclear Brinkmanship with Two-Sided Incomplete Information." *American Political Science Review* 82:155–78.

Rasmusen, Eric. 1989. *Games and Information: An Introduction to Game Theory.* New York: Basil Blackwell.

Rhodes, Edward. 1988. "Nuclear Weapons and Credibility: Deterrence Theory beyond Rationality." *Review of International Studies* 14:45–62.

Schelling, Thomas C. 1960. *The Strategy of Conflict.* New Haven: Yale University Press.

———. 1966. *Arms and Influence.* New Haven: Yale University Press.

Selten, Reinhard. 1975. "A Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory* 4:25–55.

Sobel, Joel. 1985. "A Theory of Credibility." *Review of Economic Studies* 52:557–73.

Smoke, Richard. 1987. *National Security and the Nuclear Dilemma.* Reading, MA: Addison-Wesley.

Snyder, Glenn H., and Paul Diesing. 1977. *Conflict among Nations: Bargaining, Decision Making, and System Structure in International Crises.* Princeton: Princeton University Press.

Tirole, Jean. 1988. *The Theory of Industrial Organization.* Cambridge: MIT Press.

Wagner, R. Harrison. 1988. "Reputation and the Credibility of Military Threats: Rational Choice vs. Psychology." Presented at the annual meeting of the American Political Science Association, Washington, DC.

Wilson, Robert. 1985. "Reputations in Games and Markets." In *Game-Theoretic Models of Bargaining,* ed. Alvin E. Roth. New York: Cambridge University Press.

Zagare, Frank C. 1987. *The Dynamics of Deterrence.* Chicago: University of Chicago Press.

———. 1990. "Rationality and Deterrence." *World Politics* 42:238–60.