Frank C. Zagare

The Logic of Deterrence

Abstract: This article describes the important structural characteristics of a recently developed game-theoretic model of deterrence, summarizes the major deductions drawn from it, and discusses its implications for both the theory of deterrence and the current strategic relationship of the superpowers. The model shows that a credible threat and a power advantage are neither necessary nor sufficient conditions for stable deterrence. It also suggests that, even under ideal conditions, deterrence is an intricate and fundamentally fragile relationship that rests, ultimately, upon the preferences and perceptions of key decision-makers rather than upon the nature and composition of each side's strategic arsenal.

1. Introduction

No single concept has dominated the strategic field over the past forty years as has the concept of deterrence. Yet, curiously, the theory of deterrence remains woefully underspecified. Although several classic studies in each of the three 'waves' of the deterrence literature identified by Jervis (1979) can be pointed to, no single authoritative source, no seminal work, currently exists.[1] Moreover, the field is literally strewn with a mass of disconnected and seemingly contradictory hypotheses, all purportedly deduced from a common set of assumptions (for a partial listing, see Smith 1982). To appreciate the disarray of this field of study one need only reflect on the nature of the debate in the United States in the late 1960s and the early 1970s over the development of an antiballistic missile system, or the current controversies surrounding the deployment of the MX missile and the Strategic Defense Initiative, or Star Wars, program of the Reagan Administration.

The huge gulf that separates proponents and opponents on these and related issues reflects the shaky foundation upon which the theoretical edifice of deterrence theory rests. Inexplicably, the underpinnings of deterrence theory have been more or less ignored since the early 1960s with the demise of American nuclear superiority and the 'second wave' of theorizing associated with it. Indeed, a careful inspection of this foundation reveals several faults and cracks in the underlying architecture

Analyse & Kritik 9 (1987), S. 47–61

(Zagare 1987). If deterrence were a building, it would probably be condemned.

To lay a sounder foundation for the theory of deterrence, I have elsewhere constructed a new model of deterrence and applied it to both the 1967 and 1973 crises in the Middle East, as well as to the full sweep of the strategic relationship of the United States and the Soviet Union (Zagare 1987). In my opinion, this model and the theory of moves framework that I use to analyze it (Brams 1983; Zagare 1984; Kilgour/Zagare 1986) has proven to be an extremely potent device for analyzing deterrence and for generating insights into its dynamics. My purpose in this paper is to offer a description of the important structural characteristics of this model in terms of the current relationship of the superpowers, to summarize the major deductions drawn from it, and to discuss its wider implications for the theory of deterrence.

2. A Brief Resume of a Deterrence Model

To this end consider for now the generalized representation of the superpower deterrence game depicted in Figure 1. In this game, each of the superpowers is assumed to have two strategies, one that supports the status quo (a_1 or b_1) and one that does not (a_2 or b_2). These two strategies, in turn, give rise to 2 x 2 = 4 possible outcomes, summarized verbally in Figure 1. These outcomes are represented by an ordered pair in each cell of the outcome matrix. By convention, the first entry of each pair represents the payoff to the row player (here, the United States), and the second entry the payoff to the column player (here, the Soviet Union), should that outcome be selected by the players.

		USSR: b_1	USSR: b_2
US:	a_1	STATUS QUO (a_1,b_1)	VICTORY FOR USSR (a_1,b_2)
	a_2	VICTORY FOR US (a_2,b_1)	MUTUAL LOSS (a_2,b_2)

FIGURE 1: GENERALIZED REPRESENTATION OF THE SUPERPOWER DETERRENCE GAME

For instance, if both players select their first strategy, the outcome associated with the status quo, (a_1,b_1), results. In this case, the payoff to the United States is a_1 while the payoff to the Soviet Union is b_1.

Implicit in a deterrence game such as the one depicted in Figure 1 are several assumptions about the utility functions, (u), of the two players. For example, it seems safe to assume that in any deterrence game each player would like to prevent the other player from taking some unspecified action that would upset the status quo, that is:

For the US: $u(a_1,b_1) > u(a_1,b_2)$, and (1a)
for the USSR: $u(a_1,b_1) > u(a_2,b_1)$. (1b)

Moreover, while the requirements expressed in equation (1) are necessary for a game to qualify as a deterrence game, they clearly are not sufficient. For the notion of deterrence to be of some salience, at least one player must have an incentive to move away from the status quo. Games wherein this minimal condition is satisfied will be termed unilateral deterrence games. When both players have an incentive to upset the status quo, a game of mutual deterrence will be said to exist.[2]

Unilateral deterrence games, then, by definition, meet the restrictions of equation (1) and are characterized by two players with asymmetric motivations: a status quo player who prefers the status quo to all other outcomes and a revisionist player who prefers, unilaterally, to change it. Hence, in a unilateral deterrence game where, say the United States, is postulated to be the status quo player and where, say the Soviet Union, is postulated to be the revisionist player, the following restrictions will characterize the preference orders of the two players:

For the US: $u(a_1,b_1) > u(a_2,b_1) > u(a_1,b_2)$, and (2a)
for the USSR: $u(a_1,b_2) > u(a_1,b_1) > u(a_2,b_1)$. (2b)

By contrast, in a game of mutual deterrence, the preferences of the two players with respect to these same outcomes are completely symmetric: each player prefers, unilaterally, to upset the status quo and prefers that the other player not upset it. In mutual deterrence games, then, the following restrictions on the preference orders of the two players will hold:

For the US: $u(a_2,b_1) > u(a_1,b_1) > u(a_1,b_2)$, and (3a)
for the USSR: $u(a_1,b_2) > u(a_1,b_1) > u(a_2,b_1)$. (3b)

From the above it is easy to see that once a deterrence game has been identified as either unilateral or mutual, each player's preference for three of the four outcomes are accounted for.[3] But to complete the ordering and, hence, to fully determine the nature of the deterrence game, it is necessary to specify the preference relationship of these three outcomes to (a_2,b_2). (a_2,b_2) represents the outcome that would be induced if

one player upset the status quo in order to gain a unilateral advantage and the other player resisted and attempted to punish the first and deny his opponent these advantages. Put differently, (a_2,b_2) represents the threat upon which the deterrence relationship rests.

Patently, each player's evaluation of this threat outcome is a function of the capability of the other; and each player's perception of the other's evaluation of this outcome depends upon the credibility of the other player's threat. Thus, if capability is defined as the ability to hurt (Schelling 1966), each player will have a capable threat if and only if the other player prefers that, if he takes the prohibited action, the threat not be carried out. It follows, therefore, that if both players have a capable threat:

For the US: $u(a_2,b_1) > u(a_2,b_2)$, and (4a)
for the USSR: $u(a_1,b_2) > u(a_2,b_2)$. (4b)

By contrast, if one player has a capable threat and the other does not, the preference ranking of the opponent of the player whose threat is capable would simply reverse the restriction of equation (4). And if neither player is capable of hurting the other, the preferences of both players would be reversed.

Finally, credibility. By most accounts, credibility is the 'magic ingredient' of every deterrence relationship (Freedman 1981, 96). Credibility means that the player being deterred must believe that the threat will be carried out if he takes the prohibited action. The essence of credibility, then, resides in a subjective evaluation on the part of the player being deterred of the willingness, or preference, of the other player to execute his threat. Unless the player being deterred perceives that the other prefers to resist, rather than accept, a unilateral departure from the status quo, a threat will not be seen to be credible (Fraser/Hipel 1979, 802).

Note that such an evaluation may not necessarily correspond with objective reality. A threat that is believed will be credible, whether or not the player making it intends to carry it out. Similarly, whatever the intentions of the threatener are, its threat will not be credible unless it is believed by the other player. Hence, if both players in a deterrence game have a credible threat, the following restriction will apply:

the USSR perceives that for the US: $u(a_2,b_2) > u(a_1,b_2)$, and (5a)
the US perceives that for the USSR: $u(a_2,b_2) > u(a_2,b_1)$. (5b)

As before, the lack of a credible threat by one or both players can be reflected by appropriate modifications of the direction of the inequalities in these equations.

Depending upon whether the deterrence relationship is unilateral or mutual, and whether each player's threat is capable[4] or credible, a number of structurally distinct deterrence games can be identified. In what follows, a theory of moves analysis will be used to indicate the conditions under which deterrence can be expected to succeed or fail (i.e., is stable) in each of them.

A. Mutual Deterrence. In Figure 2, the three core games of mutual deterrence are listed. In this representation, the (ordinal) payoffs of the two players are ranked from "1" to "4", with "4" representing each player's best outcome, "3" each player's next-best outcome, and so on. Thus, when this convention is adopted, the outcome (4,1) represents the best outcome for the United States and the worst outcome for the Soviet Union.

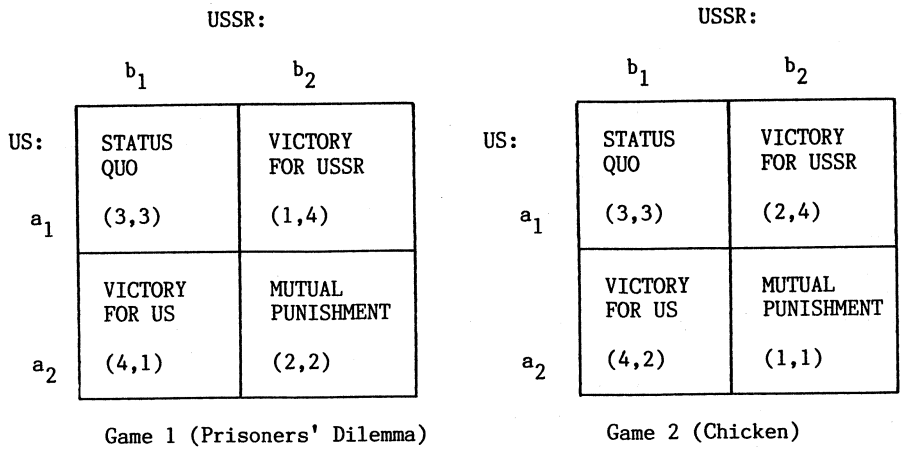
Note that each game satisfies the restrictions of equations (3a) and (3b), that is, each player is assumed to prefer, unilaterally, to upset the status quo and to prefer that the other not upset it. The three games are distinguished from one another only by different assumptions about the credibility of each player's threat.

In game 1 (Prisoners' Dilemma), both players are postulated to have a credible threat, that is, both prefer to resist rather than accept a unilateral deviation from the status quo by the other. In game 2 (Chicken), neither player has a credible threat. And in game 3 (Called Bluff), one player (in this case the United States) is assumed to have a credible threat while the other (i.e., the Soviet Union) is not.

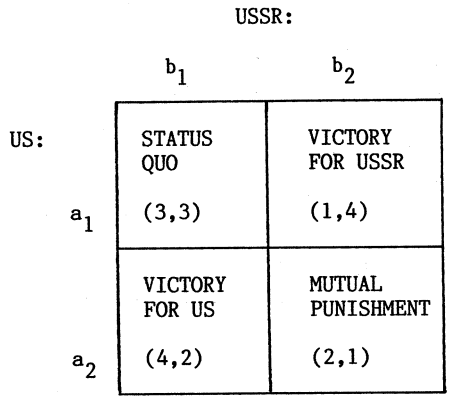
In two of these three games (1 and 2), deterrence can be stable, although the conditions that must be satisfied for stability to persist are somewhat different in each case. In the Prisoners' Dilemma game (1), wherein both players have a credible threat, deterrence is stable as long as each player has a second-strike capability, that is, the ability to move to the outcome associated with mutual punishment (i.e.,(a_2,b_2)) should the other player depart from the status quo. (For a demonstration, see Zagare 1985b). The reason for this is that, in this game, it is precisely each player's threat to induce this outcome that removes the incentive of the other player to seek a unilateral advantage by upsetting the status quo.

Interestingly, deterrence may also be stable if both players lack a credible threat. The conditions upon which deterrence stability rests in Chicken, however, are more stringent than in Prisoners' Dilemma. In Chicken, stable deterrence depends upon the ability of both players to move through mutual punishment, as in a limited war. Finally, in Called Bluff (game 3), wherein only one player has a credible threat, deterrence is not stable.

Under these conditions, the player who is willing, and able, to punish a departure from the status quo by the other wins.



Game 1 (Prisoners' Dilemma) Game 2 (Chicken)



Game 3 (Called Bluff)

FIGURE 2: THREE MUTUAL DETERRENCE GAMES

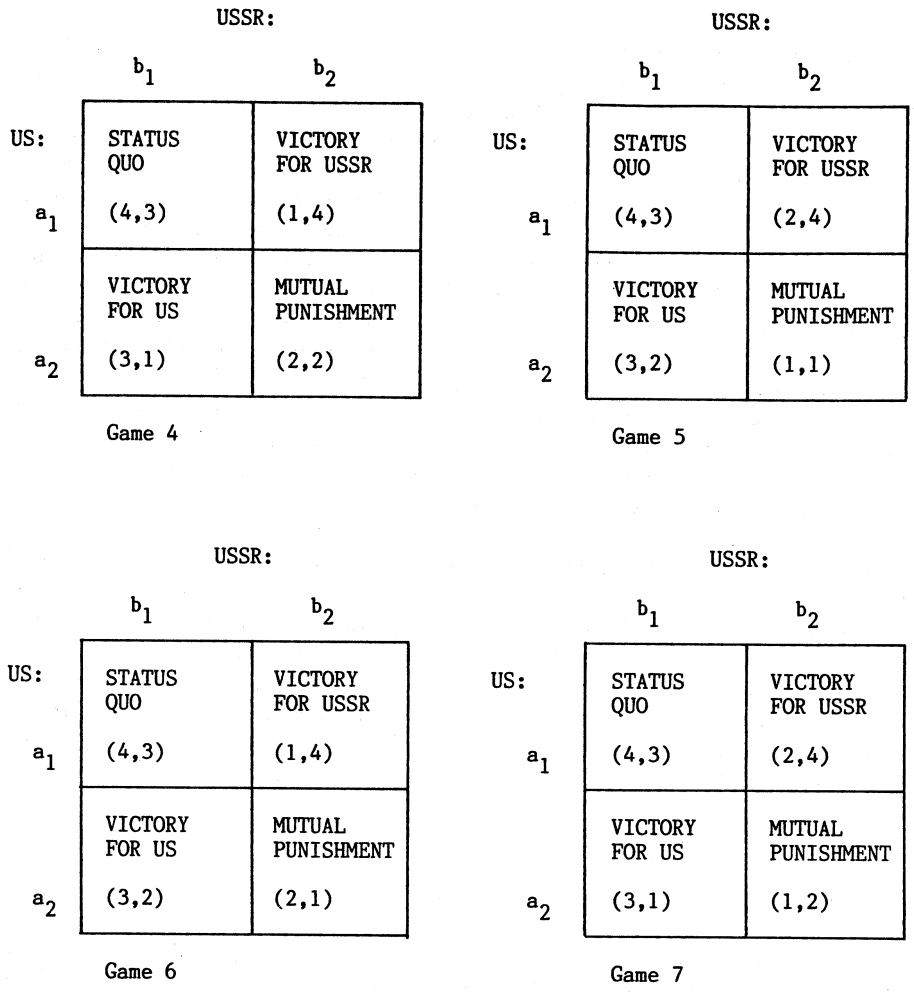


FIGURE 3: FOUR UNILATERAL DETERRENCE GAMES

These conclusions, moreover, are relatively unaffected by power asymmetries (as distinct from capabilities).[5] Only when neither player has a credible threat does a power imbalance enter into the deterrence equation. Under these conditions, deterrence is not stable since the more powerful player cannot be deterred from upsetting the status quo and inducing its best outcome as the final outcome of the game.

B. Unilateral Deterrence. Paradoxically, deterrence is much more difficult to achieve in the four core games of unilateral deterrence listed in Figure 3 than in the three games of mutual deterrence discussed above,[6] since stronger assumptions are necessary to induce a stable outcome in the unilateral deterrence games. More specifically, for one outcome to be rendered stable in each of the games of Figure 3 (significantly, it is the status quo), the revisionist player must prefer the certain selection of his next-best outcome to the lottery that contains his best and two worst outcomes.[7] As explained in greater detail in Zagare (1985a), such an assumption is more likely to be satisfied when the revisionist player is risk-averse, that is, when he deflates the (cardinal) value of his best and two worst outcomes relative to the value of his next-best outcome. Thus, the success of unilateral deterrence depends less on the credibility of each player's threat than on the attitude of the revisionist player toward risk. As long as the revisionist player is risk-averse, the status quo is stable in unilateral deterrence games. Otherwise, the games are indeterminate.

Power asymmetries, however, may upset this conclusion. In a unilateral deterrence game with an unequal distribution of power, a status quo power can ensure its best outcome and, in the process, stabilize the deterrence relationship, if it has a credible threat or the revisionist player lacks one. Still, preponderance alone is not sufficient for the success of deterrence in its unilateral variant. If a more powerful status quo player lacks credibility and the revisionist player does not, deterrence is not stable. Similarly, deterrence is also unstable when a weaker status quo player's threat is not credible.

3. General Conclusions and Implications

Tables I and II summarize the conclusions drawn from a theory of moves analysis of the strategic relationship of the United States and the Soviet Union. Each table controls for two different power configurations and the four, logically possible, assumptions that can be made about the credibility of each player's threat. Table I posits a game of mutual deterrence while Table II assumes a relationship of unilateral deterrence.

CREDIBLE THREAT	EQUAL POWER	US IS PREDOMINANT
Both Credible:	stable mutual deterrence	stable mutual deterrence
US only:	US Victory	US victory
USSR only:	USSR victory	USSR victory
Neither Credible:	stable mutual deterrence*	US victory

* Depends upon the ability of the players to pass through (a_2, b_2).

TABLE I: MUTUAL DETERRENCE

	US is a status quo power USSR is a revisionist power		US is a revisionist power USSR is a status quo power	
CREDIBLE THREAT	EQUAL POWER	US IS PREDOMINANT	EQUAL POWER	US IS PREDOMINANT
Both Credible:	stable deterrence if USSR is risk-averse; otherwise indeterminate	USSR is deterred	stable deterrence if US is risk-averse; otherwise indeterminate	US is deterred
US only:	stable deterrence if USSR is risk-averse; otherwise indeterminate	USSR is deterred	stable deterrence if US is risk-averse; otherwise indeterminate	US victory
USSR only:	stable deterrence if USSR is risk-averse; otherwise indeterminate	USSR victory	stable deterrence if US is risk-averse; otherwise indeterminate	US is deterred
Neither Credible:	stable deterrence if USSR is risk-averse; otherwise indeterminate	USSR is deterred	stable deterrence if US is risk-averse; otherwise indeterminate	US victory

TABLE II: UNILATERAL DETERRENCE

Two important qualifications about these deductions are in order. First, all of these conclusions rest upon the assumption that each player has, at minimum, a second-strike capability, that is, the ability to respond should the other player move from the status quo. (In one case more stringent assumptions are required to generate the listed finding; this exception is indicated in a table note.) And second, under certain conditions, almost all of these results can be disturbed.[8] Since both tables would be unduly complicated if every possible exception were noted, I have listed only those conclusions that are both theoretically and empirically relevant to the strategic relationship of the United States and the Soviet Union.

Tables I and II can be use to draw several interesting conclusions about both the nature of deterrence interactions in general and the dynamics of the strategic relationship of the superpowers in particular. First, note that credibility is not quite the magic ingredient that it is claimed to be, although it remains a very important part of the deterrence relationship nonetheless. Credibility is neither a necessary nor a sufficient condition for successful deterrence. For instance, in a game of unilateral deterrence where both players have a credible threat, deterrence may fail if the revisionist player is not risk-averse. Moreover, as demonstrated in Zagare (1987), without a capable threat, a player whose threat is credible will also be unable to deter an opponent who prefers to upset the status quo. By contrast, under some conditions, deterrence may constitute a stable relationship, even when each player's threat lacks inherent credibility. To wit, provided that the outcome associated with mutual punishment does not imply termination of the game, and provided that neither player possesses clear-cut conventional superiority, mutual deterrence remains stable even when both players have incredible threats. Deterrence is also stable under these conditions in unilateral deterrence situations if the balance of power favors the status quo player.

Like credibility, a power advantage is neither necessary nor sufficient for deterrence to work. A weaker player with a capable and credible threat, for instance, should be able to deter a stronger opponent. Conversely, a dominant power may, under certain conditions, be unable to deter its weaker adversary. Thus, at the theoretical level at least, deterrence stability is not a direct consequence of either a balance or an imbalance of power.[9]

All of which suggests that deterrence interactions are intricate and unusually complex. There is no one-to-one relationship between any single dimension of the model and overall deterrence stability. Slight alterations in one or another parameter, such as the nature of each player's retaliatory threat, power, offensive ability, weapons characteristics, and attitude toward risk may have dramatic consequences for the success, or the failure, of deterrence. (See Zagare 1987, for additional details.)

Another interesting insight that emerges from the theoretical development of the model stems from the structural implications associated with game theory's notorious Prisoners' Dilemma game. If this game in fact represents the mutual deterrence relationship in its ideal manifestation - that is, when both players have a capable and credible retaliatory threat - then it follows that players in deterrence and related crisis games will have an incentive to try to create or reinforce interactions that share its structure. This inference is supported in several empirical applications of the model. For instance, in both the 1967 and the 1973 crises in the Middle East, decision-makers in the Soviet Union and the United States exhibited a tendency to transmit tit-for-tat communications to one another, conveying both a willingness to compromise, but also a determination to respond to the untoward behavior of the other. Thus, not only does the structure of this game describe the conditions conducive to stable mutual deterrence, but it also explains the obvious and well-documented discrepancies between the behavior exhibited by states during actual crisis situations and some of the more esoteric crisis management strategies suggested by some deterrence theorists (Young 1968; Snyder/Diesing 1977).

The stability characteristics of the status quo in Prisoners' Dilemma are also suggestive of another salient characteristic of these relationships. Deterrence constitutes a stable relationship in this game only when the compromise outcome is the status quo and both players have the ability to punish departures from it. In the absence of these conditions, deterrence is unstable and conflict is implied. Thus, even in its ideal form, deterrence is rickety and fragile.

The delicate balance of terror is underscored by a historical analysis of the strategic relationship of the United States and the Soviet Union. Fortunately, this relationship has remained stable despite the fact that it has undergone considerable evolution since its inception in 1945. Some may interpret the continuing stability of the superpower relationship as evidence of the robustness of deterrence. But since some rather disturbing assumptions are required to explain the observed stability of certain periods of the post-war era, there is more here than meets the eye. For instance, if one argues that only the United States was interested in upsetting the status quo from 1946 to 1948, as does Howard (1983), or if one argues that only the Soviets were revisionist during this period, as do Intriligator and Brito (1984, 82) or Brodie (1959), then Soviet risk-aversion - in the first case - or American risk-aversion - in the second case - must be assumed to explain the absence of a superpower war at this time. Similarly, one cannot explain stability from 1962 to 1966 when the United States possessed a first-strike capability (Quester 1970, 216; Friedberg 1982, 69) without also assuming that the United States was a status quo power. Finally, one cannot explain the absence of a superpower war since 1967 without also assuming that at least one superpower was able to

make credible a threat to retaliate at a time when retaliation implied the destruction of its political, economic and social system. The problematic nature of this assumption is underscored in Achen's (1986) latest contribution to the literature.

It appears, then, that deterrence is not directly a function of those variables that are most easily manipulated by decision-makers, that is, the nature and composition of each side's strategic arsenal. Rather deterrence stability resides 'in the heads' of world leaders. It depends not only on obvious objective factors, but may also hinge on the preferences and perceptions (or misperceptions) of those who have the ability to induce Armageddon. Such factors, unlike the balance of military power, can be subject to quick and erratic changes, as coup d'états, revolutions, elections, illness, and other forces bring about fundamental leadership changes. Deterrence is indeed a tenuous relationship.

Given the fragile nature of the deterrence relationship, one is naturally led to ask if there is a better way to manage world affairs. I have no magic elixir to offer here. Perhaps, some day, other approaches to international relations will render unnecessary such a ghastly approach to the world about us, provided that deterrence can deliver us to the millenium. Until then, however, we surely remain 'prisoners' of our own imagination.

Notes

1 According to Jervis, the basic concepts and the fundamental assumptions of deterrence theory were articulated in the first 'wave' of the deterrence literature immediately after World War II by writers such as Bernard Brodie and Arnold Wolfers. About ten years later, a second wave of theorizing occurred that revived and extended the literature of the first wave. Jervis includes writers such as Thomas Schelling and Glenn Snyder in this category. The third and most recent wave of the literature is characterized by an attempt to subject the propositions developed during the first two waves to empirical scrutiny. (For a more detailed discussion of this characterization of the field see Smith 1982, 23-28.)

2 This is a departure from conventional usage. These terms are normally used to indicate a deterrence relationship wherein either one player (unilateral deterrence) or both players (mutual deterrence) have a nuclear capability. For a discussion, see Morgan (1983).

3 The rankings expressed in equations (2) and (3) reflect the additional assumption that each player prefers the outcome associated with his own departure from the status quo to the outcome implied by his opponent's departure from it. For a justification of this assumption, see Zagare (1985a).

4 To simplify the following exposition, each player will be assumed to have a capable threat. In the nuclear age, such an assumption seems warranted.

5 By power I mean "the ability to prevail in conflict" (Deutsch 1978, 23). Although several distinct notions of power have been developed within the theory of moves framework, I have used the concept of holding power (Kilgour/Zagare 1986) to measure the impact of an asymmetric distribution of power in a deterrence game. Holding power is simply the ability of one player in a sequential game to stay at an outcome longer than his opponent.

6 For the sake of convenient exposition, the Soviet Union is listed as the revisionist player in Figure 3. No value judgement is implied. In the subsequent discussion, the ramifications of the opposite assumption will be explored.

7 By contrast, for an outcome to be rendered stable in the mutual deterrence games of Figure 2, the players need only be able to look ahead and consider the long-term consequences of departing from the status quo.

8 For the particulars of the formal deductions, the reader is referred to Zagare (1987).

9 In Zagare (1987, chapter 4), however, I demonstrate that capability is a necessary condition for deterrence stability. In other words, deterrence will always fail when a revisionist player's opponent lacks capability.

Bibliography

Achen, C.H. (1986), A Darwinian View of Deterrence, in: J. Kugler/F.C. Zagare (eds.), The Stability of Deterrence. Monograph Series in World Affairs, Denver

Brams, S.J. (1983), Superior Beings: If They Exist, How Would We Know? New York

Brodie, B. (1959), The Anatomy of Deterrence, in: World Politics 11, 173-179

Deutsch, K.W. (1978), The Analysis of International Relations, 2nd ed., Englewood Cliffs/NJ

Fraser, N.M./K. Hipel (1979), Solving Complex Conflicts, in: IEEE Transactions on Systems, Man, and Cybernetics, SCM-9, 12, 805-16

Friedberg, A.L. (1982), The Evolution of U.S. Strategic 'Doctrine' - 1945 to 1981, in: S.P. Huntington (ed.), The Strategic Imperative: New Politics for American Security, Cambridge

Freedman, L. (1981), The Evolution of Nuclear Strategy, New York

Howard, M. (1983), The Causes of War, Cambridge

Intriligator, M.D./D. L. Brito (1984), Can Arms Races Lead to the Outbreak of War? in: Journal of Conflict Resolution 28, 63-84

Jervis, R. (1979), Deterrence Theory Revisited, in: World Politics 31, 289-324

Kilgour, D.M./F.C. Zagare (1986), Holding Power in Sequential Games, in: International Interactions 13, 159-182

Morgan, P.M. (1983), Deterrence: A Conceptual Analysis, 2nd ed., Beverly Hills

Quester, G.H. (1970), Nuclear Diplomacy, New York

Schelling, T.C. (1966), Arms and Influence, New Haven

Smith, T.C. (1982), Trojan Peace: Some Deterrence Propositions Tested, Monograph Series in World Affairs, Denver

Snyder, G.H./P. Diesing (1977), Conflict Among Nations: Bargaining, Decision Making and System Structure in International Crises, Princeton NJ

Young, O.R. (1968), The Politics of Force: Bargaining During International Crises, Princeton/NJ

Zagare, F.C. (1984b), Limited-Move Equilibria in 2 x 2 Games, in: Theory and Decision 16, 1-19

- (1985a), The Pathologies of Unilateral Deterrence, in: U.Luterbacher/M.D. Ward (eds.), Dynamic Models of International Conflict, Boulder CO

- (1985b), Toward A Reformulation of the Theory of Mutual Deterrence, in: International Studies Quarterly 29, 155-169

- (1987), The Dynamics of Deterrence, Chicago