O'Neill, Barry

Working Paper

A Measure for Crisis Instability

Discussion Paper, No. 652

Provided in Cooperation with:
Kellogg School of Management - Center for Mathematical Studies in Economics and Management Science, Northwestern University

Suggested Citation: O'Neill, Barry (1985) : A Measure for Crisis Instability, Discussion Paper, No. 652, Northwestern University, Kellogg School of Management, Center for Mathematical Studies in Economics and Management Science, Evanston, IL

This Version is available at:
https://hdl.handle.net/10419/221011

Discussion Paper No. 652

A MEASURE FOR CRISIS INSTABILITY

by

Barry O'Neill

March 1985

Department of Industrial Engineering
and Management Sciences
Northwestern University
Evanston, Illinois 60201

Abstract

An index is described to measure crisis instability, the danger of pre-emptive war due to each government's fear that the other is about to attack. Crisis instability is interpreted as the probability of war an observer should assess knowing only the costs to each government for striking first and for being attacked. A series of axioms defines a Crisis Instability Index, a scale of danger that gives a unique rank ordering of crisis situations. The measure is equivalent to a simple function of the critical risks of the two governments, as defined by Ellsberg.

A companion paper discusses the stability consequences of various arms control agreements and of building anti-missile weapons in space.

1. Introduction

A widely-expressed hope is that the nuclear deterrent system is safe as long as sane leaders are in control, since any attack would bring on a devastating reprisal. The implicit assumption here is that the attacking government faces a clear choice of war or peace, and so would choose peace. However, the nuclear threat system allows other ways a war might start in which no leader would face the simple choice of peace over war.

One way is the accidental, unauthorized or third-party use of nuclear weapons. A second is escalation, as governments raise the risk of ultimate nuclear war to achieve their ends in small conflicts. A third, which we will analyze here, is the phenomenon of crisis instability.

The Concept of Crisis Instability

During a crisis, when antagonists have invoked the nuclear threat implicitly or explicitly, each government may feel compelled to attack first, not because it prefers war to continued peace but because it fears an attack and wants to lessen the catastrophic effects to itself. Each government knows the other does not want war, but suspicion grows that the other is planning an attack or perhaps only that the other is worrying about being attacked, leading one side to "retaliate first."

A 2400-year-old source, Xenophon's history of the Persian Expedition (1949, p.82, quoted by Schelling, 1984), states the logic clearly. Tension arose between the Greek army departing Persia and the Persian force escorting it, so the Greek leader addressed his Persian counterpart as follows:

> "I observe you are watching our moves as though we were enemies, and we, noticing this, are watching yours too. I know that cases have occurred in the past when people, sometimes as the result of suspicion have become frightened of each other and then in their

> anxiety to strike first before anything is done to them, have done irreparable harm to those who neither intended nor even wanted to do them any harm at all. I have come then in the conviction that misunderstandings of this sort can best be ended by personal contact, and I want to make it clear to you that you have no reason to distrust us."

The Greek general recognized the two elements that destroy stability: the existence of an objective situation that gives an advantage to the first mover, and event of psychological impact that initiates the spiral of mutual suspicion that the other is about to strike. The importance of this catalyst event justifies the name crisis instability, although specifying what types of events can trigger the fear of war seems to be a difficult problem. The more predictable feature is the degree of offensive advantage of the weapons holdings. Just as arms race theory holds that weapons themselves cause more weapons, crisis instability theory regards weapons as intrinsic causes of war.[1]

The current nuclear threat system has many elements that give the advantage to the offense. One is the continuing rise in missile accuracy. Another is the introduction of multiple warheads (MIRVs) which allow each missile to destroy several of the opponent's missiles in their silos, each of the latter missiles having several warheads, so that ICBMs with MIRVs give a double incentive to play the role of attacker. Failure to find an invulnerable basing mode for land-based missiles, the possibility of attacking the other's command and control system, the prospect of anti-submarine warfare, and the vulnerability of the complex systems now proposed for ballistic missile defense may increase the attractiveness of striking first.

Past Theories of Crisis Instability

Concern over crisis instability has influenced American and Soviet arms control policy since the late 1960's, but analysts have disagreed about what

the concept means, and whether particular systems help or hurt it. Some state that the "essential component" of crisis stability is invulnerability (e.g., Gray, 1981), others emphasize the number of each side's warheads per opponent's silo, and others state that stability amounts to the effectiveness of the retaliatory strike (Payne, 1984). The discrepancy does not stem from any disagreement about objective properties of weapons systems but from a vagueness in the concept and its surrounding logic. A number of authors have listed factors that influence crisis instability but do not say just how these determine it. Usually the factors trade off with one another, so unless we know their functional relationship to instability, we cannot draw conclusions for policy.

One way to put crisis instability on a more solid foundation is to analyse it formally. Ellsberg (1961) was the first to treat it mathematically, and Hunter (1972) and Wagner (1983) have extended his theory. Kupperman and Smith (1978), Hughes (1978) and Grotte (1982) have provided other mathematical approaches.

Ellsberg calculated each side's critical risk, the largest risk acceptable to a government that the other is about to strike, "acceptable risk" meaning that a government should choose to wait when its assessed probability of the other striking lies below that value. He postulated that the larger this risk, the more stable the situation.

There are two difficulties with this approach. First, it generates two indices of temptation to strike, one for each government, but does not tell us how to combine them to get a single measure of the danger of the whole situation. If one side altered its weapons holdings in a way that raised its critical risk but lowered the other's, would this represent an overall improvement?

Secondly, the degree of danger should depend not on each side's critical probability of the other striking, but on the relative values of the critical and the assessed probabilities, i.e., the probability the side really holds that the other is about to strike. If I see an antagonistic government introducing a vulnerable multi-warhead ICBM my critical risk might rise (since its retaliation would be costlier to me), but so might my assessed risk (since I know it fears the loss of its weapons if it lets me strike first). The root problem with Ellsberg's analysis in our view is that a side's critical risk involves its own costs but not other side's costs. The decision should depend on all costs: my tendency to strike first should grow with my belief that you are contemplating a first strike, which should depend on my perceptions of how painful such a move would be to you.

The second section of this paper describes a development from Ellsberg's theory and from Grotte's related work (1980). It defines a single numerical index that measures crisis instability based on the outcomes to each government from attacking first and from being attacked. In a further paper (O'Neill, 1985) we will describe a model for the outcomes of a hypothetical nuclear war, which will allow us to estimate the stability effects of various arms control measures and of space-based anti-missile systems.

2. The Model of a Crisis

We will make a series of assumptions about the dynamics of a crisis. One could add further complexities to the model based on specific beliefs about the dynamics of certain types of crises and follow the lines described here to derive a measure of crisis instability, but the assumptions given now are simple and keep the core structure of the dilemma

the decision-makers would face.

A crisis is assumed to occur in one stage during which each government makes a decision whether or not to strike. If each decides not to, the crisis is over. To rationalize this ending rule one could assume that the objective situation has changed or that the news that the other side decided not to strike when it had the chance has convinced each government that it will not attack now.

We assume that if a crisis occurs one government has the opportunity to strike first. One way to interpret this assumption is that one of the two governments is going to have the information sooner or have a faster response time. The governments are equally likely to have this chance and neither knows whether it possesses the first or second move. That is to say, if it tries to strike first it does not know whether it will actually succeed in being first or whether the other is in position to achieve a pre-emptive attack.

FIGURE 1 HERE

The extensive-form game of Figure 1 represents this situation. In Figure 1 the chance move at the beginning determines which government can act first, but the game is formally equivalent to one in which the chance event occurs at the end and where each player decides on a <u>disposition</u> of how to behave, with a random event selecting who is able to strike first if both have decided to do so. The latter game has a different temporal order of events than occurs in the world, but is strategically identical to the former game and has the advantage that it can be represented by a matrix (Matrix 1).

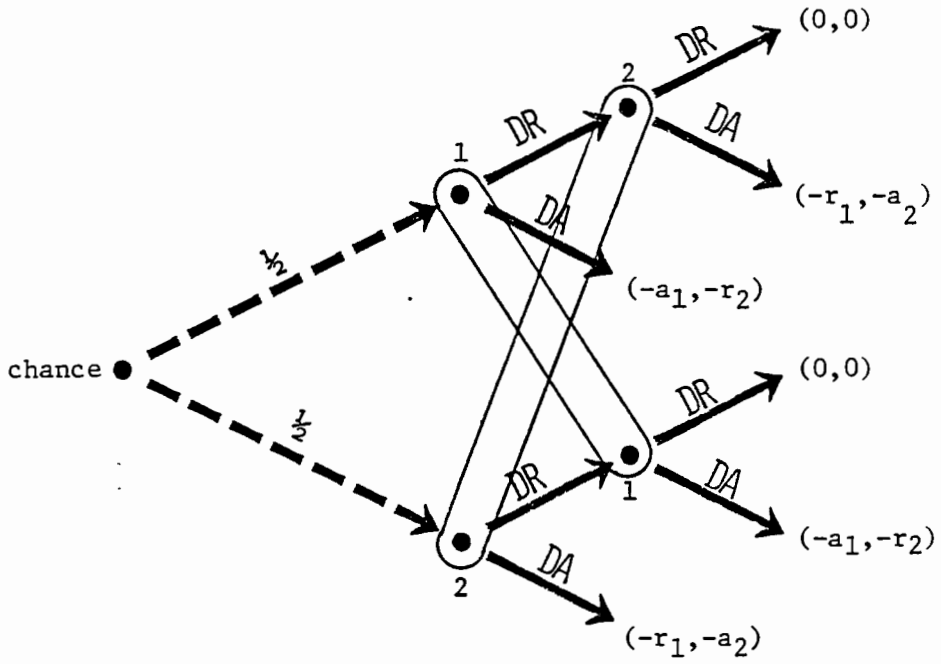

Figure 1: The extensive-form game model of a crisis.
DA: decides to attack; DR: decides to refrain from attacking. Loops indicate that the player does not know which position the play has reached.

		Gov't 2: Decide to Refrain	Gov't 2: Decide to Attack
Gov't 1	Decide to Refrain	$0, 0$	$-r_1, -a_2$
	Decide to Attack	$-a_1, -r_2$	$\frac{-a_1-r_1}{2}, \frac{-a_2-r_2}{2}$

Matrix 1. The crisis as a game matrix.

The values $-a_i$ and $-r_i$ represent the utilities of the government i when it attacks first or attacks second respectively and satisfy $-r_i < -a_i < 0$. The zero point on the two utility scales has been arbitrarily set as the utility of peace. An example of such a game is given in Matrix 2.

		Gov't 2: Decide to Refrain	Gov't 2: Decide to Attack
Gov't 1	Decide to Refrain	0,0	-45,-12
	Decide to Attack	-15,-20	-30,-16

Matrix 2. An numerical example of Game Type 61

Note that both prefer peace over any other outcome, but would prefer to strike first rather than be struck. The fourth cell is the average of the two other war cells, reflecting their equal likelihood of achieving a first strike if both try to do that. "Type 61" is the designation for games with this ordering of payoffs according to the classification scheme of Rapoport and Guyer (1966), while Jervis (1978) calls it and its n-person

generalization the "Stag Hunt".

The outcome (Refrain, Refrain) is most attractive because it is payoff dominant -- it is better for both players than any other outcome in the game. However rational players might not choose it, especially with poor communication and a history of conflict. The reason is the essence of the concept of crisis instability: neither can afford to risk restraint. The more extreme example of Type 61 in Matrix 3 illustrates the point that an outcome may be best for all but not be the players' choice. Even though peace at (0,0) is the payoff dominant outcome, the players do not dare to trust one another, and (-2,-2) seems to be the likely result.

		Gov't 2: Decide to Refrain	Gov't 2: Decide to Attack
Gov't 1	Decide to Refrain	0,0	-1000,-1
	Decide to Attack	-1,-1000	-2,-2

Matrix 3. An extreme example of Game Type 61.

There are three equilibrium strategy pairs in Matrix 1: Peace (Refrain, Refrain), Mutual Attack (Attack, Attack), and a third equilibrium involving mixed strategies. We will investigate only the first two because they are strong equilibria, meaning that any deviation from them lowers that player's utilities. (In contrast a lone deviation from the mixed strategy equilibrium gives that player the same payoff as the use of the equilibrium strategy). Also mixed strategies seem unrealistic for use by governments: almost by definition leaders want to have control, and certainly they do not want to make crucial decisions using a random device.

We assume that players will choose one equilibrium or the other, that

they will not end up with one having chosen Refrain and the other Attack. We are assuming in effect that if a crisis occurs, various factors, some game-theoretic and some not, will point to one of the two moves as being the other player's likely choice. Given the structure of the game with its strong equilibria, a player with enough valid information about the other's incipient choice, will have an incentive to make the same choice, so we are assuming that expectations will be strong and valid.

This assumption means that our analysis does not cover situations where one side gets misinformation that the other is about to attack and strikes first to the complete surprise of the other. Instead our model assumes common knowledge about world tension.

3. Axioms for the Crisis Instability Index

The crisis instability of a matrix will be defined as the likelihood of the war equilibrium assessed by an observer who is rational but limited in knowledge, who knows only the utilities to each government for striking first or second.[2] Since the observer makes the judgment before the details of the crisis are known, the likelihood will not be influenced by such other factors as the issue of the crisis, efforts to control it, or any events at the time of the crisis hinting at peace or war. It will depend only on the possible war outcomes as given in Matrix 1.

The strong equilibrium outcomes are labeled a, b, c, etc. The function $P(a|G)$ gives the rational subjective probability of equilibrium a, given that the players are in the game G. We will write $P(a|G)$ as $P_G(a)$. For games with three equilibria we use the notation $P_G(a|a \text{ or } b)$ for the likelihood of equilibrium a given that either a or b are chosen in game G.

Equilibrium probabilities for a certain class of three-person games which are not models of crisis stability will be considered first, since it

will turn out that the probabilities of equilibria in these games have implications for games of Type 61. A unanimity game has positive payoffs along the diagonal and zeros off-diagonal. Clearly an nxn unanimity game has n strong equilibria at each diagonal outcome. Unanimity games can be described by listing their equilibrium outcomes since all other payoffs are understood to be zero, and we use the notation <ab...> for a typical unanimity game, where a, b, etc., are not numbers, but possible events.

The axioms will describe the likelihoods of the equilibria in unanimity games, and link these with games of Type 61. The probability function P will apply to the set G, defined as including all 2x2 games with two strong equilibria, and all 3x3 games with three strong equilibria,

Axiom 1. (Probability) For any $G \in \underline{G}$, P_G is a conditional probability measure on the strong equilibria of G.

Axiom 1 says that a strong equilibrium is sure to be chosen and that the observer's likelihoods for the equilibria follow the standard probability axioms.

Axiom 2. (Dependence on the abstract game) The probability function P_G depends only the relationship of pairs of moves to utilities of outcomes, i.e., it is independent of the labeling or ordering of the players or moves. It is also independent of positive linear transformations of a player's utilities.

The latter part of Axiom 2 assumes that players cannot make interpersonal comparisons of utility.

Axiom 3. (Continuity/monotonicity) If G is a unanimity game with a one of its diagonal outcomes, then $P_G(a)$ is continuous and strictly monotonic in the payoffs at a.

Axiom 3 says that the more desirable an outcome for either player, the more likely it is. The axiom is appropriate for unanimity games, where the players face the same risk trying to get one payoff as another, since the off-diagonal entries are all zero. It implies that even a payoff-

dominated outcome of a unanimity game will be chosen with some positive probability, a claim that seems plausible since non-game-theoretical factors like tradition may point to such an equilibrium, with the players afraid to deviate lest they end up off the diagonal.

Axiom 4. (Probabilistic independence of irrelevant alternatives). Equilibrium probabilities for the 2x2 unanimity game <ab> and the 3x3 unanimity game <abc> are related as follows:

$$P_{\langle ab\rangle}(a) = P_{\langle abc\rangle}(a \mid a \text{ or } b).$$

This axiom declares that if the observer is informed that outcome c will not be chosen in the game <abc>, then the assessed likelihoods are identical to those in two-person game generated by eliminating c. Whether the non-choice of c is given by the rules of the game or by augmenting the observer's information is not relevant to the observer's assessment in any way that is obvious to us, so we will assume specifically that it is irrelevant. This axiom is the probabilistic analogue of Nash's independence of irrelevant alternatives axiom in bargaining theory (Nash, 1953). Luce (1959a, 1978) thoroughly investigated it in the context of single-person decision-making, and it is sometimes called Luce's Choice Axiom.

Note that by Axiom 3 the latter conditional probability will always be defined.

The next axiom makes P dependent only on those features of a game embodied in its best reply functions, a concept that was developed essentially by Harsanyi and Selten (1980, Ch.2). According to decision theory a player's choice of a move is determined by the utilities in the matrix, along with that player's assessment of the other's likelihoods of making each move. In Matrix 2 for example, if 2's subjective probability

of 1 choosing Attack is greater than .25, 2 should choose Attack. We can construct two best reply functions, one for each player, showing for every subjective probability distribution over the other's moves, which moves the player should choose. The best reply functions for Matrix 2 are shown in Figure 2.

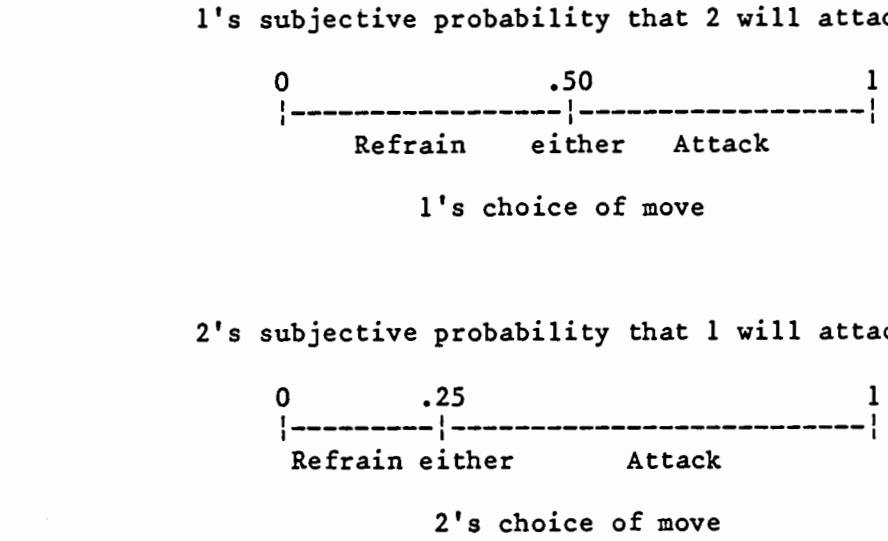

FIGURE 2. Best reply functions for Matrix 2.

A player's function is constructed entirely from the payoff matrix and, given a subjective probability distribution over the other's moves, it gives sufficient information for a choice of a move. It is conceivable that the payoff matrix contains information that should influence the probability distribution and that is not in the best reply functions, but it is not obvious at all what this extra information is. Since only the best reply functions are clearly relevant to the players, we state that they alone are relevant to the observer's assessment of likelihoods.

It is easy to show that if a constant is added to the row-chooser's payoffs in a single column, or to the column-chooser's payoffs in a single row, the resulting game will have the same best reply function. Also, in the context of 2x2 games, any two games with the same best reply functions

can be generated from one another using these operations.

Axiom 5. (Dependence on best reply functions) If G and G' have the same best reply functions, then for any strong equilibrium outcome a in G and corresponding outcome a' in G', $P_G(a) = P_{G'}(a')$.

We can now derive a general formula for the likelihood of war:

Theorem 1. For any P satisfying Axioms 1-5, there exists a $g > 0$, such that for any game G of Type 61,

$$P_G(\text{War}) = 1/\{1 + [4a_1a_2/(r_1-a_1)(r_2-a_2)]^g\}.$$

(Proofs of this and the next two theorems appear in the appendix.)

This expression contains the parameter g which corresponds to the observer's responsiveness to the game utilities as compared to the other non-payoff factors. Low values of g are associated with judgments where the unknown non-payoff factors are thought to be important so the likelihood of war is set near 50% regardless of the payoffs, and high values of g tend to place the probability at more extreme values based on the matrix alone.

We do not know how to determine a single appropriate g, but the next theorem states that any value of g gives the same rank order of the games by degree of danger as any other. Thus if we are satisfied with an ordinal measure of crisis instability, we can select g arbitrarily.

Definition: A group of measures for a set of games are equivalent ordinal scales if all rank-order the games identically.

Referring to the notation of Matrix 1 and interpreting the probability of peace as a measure of crisis instability,

Theorem 2. If games of Type 61 are measured by their probability of war, the measures generated by probability functions satisfying Axioms 1 to 5 are equivalent ordinal scales, and all are ordinally equivalent to the following measure:

$$(r_1/a_1-1)(r_2/a_2-1)$$

Definition: The Crisis Instability Index (CII) of a game of Type 61 is defined by the formula in Theorem 2.

The CII is not itself a probability as it varies between 0 and infinity, but it exhibits the directional behaviour expected of a crisis instability measure. In the limit as a_1 approaches 0, government 1 has nothing to lose by attacking first compared to not attacking at all, war becomes certain, and CII approaches infinity. Similarly as r_1 goes to a_1, CII goes to 0 -- if one government can do no better striking first than second, peace is certain.

Adding a large constant cost to each war cost a_i and r_i is equivalent to having a_i approach r_i, and increases the likelihood of peace. This event would correspond to each side acquiring the belief that a threshold effect for a nuclear winter would occur.

4. The Crisis Instability Index and Critical Risk

Following Ellsberg, for two governments in the situation of Matrix 2, we can calculate the critical risks, $\underline{q}_1$ and $\underline{q}_2$, the greatest probabilities they would assess that the other is about to strike before they themselves decide to strike. From these probabilities, we calculate the critical odds for the event the other side is planning a strike, $\underline{o}_1 = \underline{q}_1/(1-\underline{q}_1)$, and $\underline{o}_2 = \underline{q}_2/(1-\underline{q}_2)$. The higher the critical odds the better, e.g., a critical risk of .75 gives a critical odds of 3, so a government would have to hold 3 to 1 odds that the other is about to attack before it would be willing to strike first.[3]

Theorem 3. The inverse of the product of the critical odds $1/\underline{o}_1\underline{o}_2$ is ordinally equivalent to the Crisis Instability Index.

Ellsberg's rationale for his theory was quite different from our series of axioms, but surprisingly our measure of crisis instability is

equivalent to a simple function of the critical odds.

5. Discussion

The Crisis Instability Index is relatively simple and gives intuitively acceptable results when applied to specific situations as is shown in the companion paper (O'Neill, 1985).

The index has an unusual aspect that is both a strength and a weakness. It is not a theory of how governments would behave, but of how an observer knowing only the costs of war should assess their likely behaviour. It was generated by declaring non-decision matrix variables to be irrelevant to the assessment on the grounds that the observer did not know their values. The theme of the axioms was that if two possible situations were alike except for these unknown aspects, the observer should regard the two as equally likely.

A present-day specialist is in a situation much like the observer. The CII would be the rational way to estimate the benefits of a change in weapons holdings, until a time when a more comprehensive theory of crisis behaviour is developed.

The above approach, discussing the consequences of a change in one variable while assuming other variables constant, is frequently found in the non-mathematical literature of international relations. Also common in the informal literature is the opposite method: researching a past crisis situation as extensively as possible to identify all the important variables that influenced behaviour. A contribution of the Crisis Instability Index might be that the axioms would organize historical information about crisis behaviour. One can ask how each axiom was violated in real life. How did goverments assess the likelihood of the other's move? Which non-decision-matrix considerations were most important?

References

Clausewitz, C. On War. J.J. Graham (trans.), A. Rapoport (ed.). Baltimore: Penguin. 1976.

Ellsberg, D. The Crude Analysis of Strategic Choices. Rand P-2183. Santa Monica, Ca. 1960. abridged version in American Economic Review. 51 472-478, 1961.

Gray, C. Crisis stability, Ch.10 in The MX ICBM and National Security. Praeger: New York, 1981.

Grotte, J.H. Measuring strategic stability with a two-sided nuclear exchange model. Journal of Conflict Resolution, 24, 213-239, 1980.

Harsanyi, J., and R. Selten, A Noncooperative Solution Concept with Cooperative Applications. Ch.2. (mimeo). Center for Research in Management. University of California, Berkeley. 1980.

Hughes, P. Arms control and strategic stability. Air Force Magazine. 61, 1978, 56-64.

Hunter, D.E. Some aspects of a decision-making model in nuclear deterrence theory. Journal of Peace Research. 1972.

Jervis, R. Cooperation under the security dilemma. World Politics 30 167-214, 1978.

Kupperman, R., and H. Smith. A catastrophe surface in mutual deterrence theory: arms limitation and crisis stability. Journal of Peace Science. 3, 111-121, 1978.

Levy, J.S. The offensive/defensive balance of military technology: a theoretical and historical analysis. International Studies Quarterly 28, 219-238, 1984.

Luce, R.D. Individual Choice Behavior: a Theoretical Analysis. New York: Wiley, 1959a.

Luce, R.D. On the possible psychophysical laws. Psychological Review 66, 81-95, 1959b.

Luce, R.D. Comments on Rozeboom's criticism of "On the possible psychophysical laws". Psychological Review 69, 548-551, 1962.

Luce, R.D. The choice axiom after twenty years. Journal of Mathematical Psychology. 15, 215-233, 1978.

Nash, J.F. Two-person cooperative games. Econometrica 21, 129-140, 1953.

O'Neill, B. Applications of a crisis stability index: arms control agreements and space-based missile defenses. Northwestern University, mimeo. 1985.

Payne, K. Strategic defense and stability .Orbis. 216-229, 1984.

Rapoport, A., and M. Guyer. A taxonomy of 2x2 games. General Systems, 11, 203-204, 1966.

Rozeboom, W. The untenability of Luce's Principle. Psychological Review, 69, 542-547, 1962.

Schelling, T.C. The Strategy of Conflict. New York: Oxford University Press. 1960.

Schelling, T.C. Arms and Influence. New Haven: Yale University Press. 1966.

Schelling, T.C. Confidence in crisis. International Security 8, 55-66, 1984.

Snyder, G. Deterrence and Defense. Princeton: Princeton University Press. 1961.

van Damme, E. Refinements of the Nash Equilibrium Concept. Ph.D. dissertation, Technological Institute, Delft, Netherlands. 1983.

Wagner, R.H. The theory of games and the problem of international cooperation. American Political Science Review 77, 330-346, 1983.

Xenophon, The Persian Expedition. trans. Rex Warner. New York: Penguin. 1949.

Footnotes

1: Some other analysts of crisis instability since the time of Xenophon are Clausewitz (1976, p.293, quoted in Levy's comprehensive historical summary, 1984), who hoped that superiority of the defense over the offense might "tame the elementary impetuousity of war", and Schelling (1960, 1966) who referred to the "reciprocal fear of surprise attack" and the "dynamics of mutual alarm". Snyder (1961) discussed the determinants of crisis instability and expressed a belief which later developments would prove ironical, that silo-based missiles would solve the problem, as one missile could at best destroy only one of the opponent's. Jervis (1978) considered different types of crisis instability using historical examples.

Another concept of crisis instability which we do not discuss in this paper is the tendency of decisions to become worse because of the time pressure of a crisis.

2: Our approach differs from other equilibrium selection theories in that it assigns a probability to each equilibrium. Van Damme (1983) surveys the equilibrium selection problem, and all the theories he cites are deterministic in that they try to choose a single equilibrium or set of equilibria. As the game parameters change, equilibria jump into and out of the set discontinuously. Theories like this are too strong for our purposes, since we recognize that some non-game-theoretical factors will influence the player's judgements about which equilibrium to choose, and abruptly changing an equilibrium from an impossibility to a certainty based only on the utilities leaves no room for these factors. Using such a theory would imply that up to a limit adding warheads would create no danger at all, but beyond that war would be certain.

Our theory can be regarded as a probabilistic generalization of Harsanyi and Selten's equilibrium selection method for 2x2 games with two strong equilibrium (1980, Ch.2), which essentially uses Axioms 2, 3 and 5.

3: One difference between Ellsberg's and our own approaches is that he did not include a fourth outcome representing decisions by both to attack. In our model two sides can see themselves heading to war and decide to attack, but his conception was that a mutual decision to attack was negligibly likely -- an attack occurs by surprise as each side's attitudes and rationality vary over time in ways unknown to the other side. In the analysis here we have stretched his model in our direction by adding a fourth pair of payoffs to his matrix, calculated as the averages of the other two war payoffs as was done for Matrix 2.

Footnotes

1: Some other analysts of crisis instability since the time of Xenophon are Clausewitz (1976, p.293, quoted in Levy's comprehensive historical summary, 1984), who hoped that superiority of the defense over the offense might "tame the elementary impetuousity of war", and Schelling (1960, 1966) who referred to the "reciprocal fear of surprise attack" and the "dynamics of mutual alarm". Snyder (1961) discussed the determinants of crisis instability and expressed a belief which later developments would prove ironical, that silo-based missiles would solve the problem, as one missile could at best destroy only one of the opponent's. Jervis (1978) considered different types of crisis instability using historical examples.

Another concept of crisis instability which we do not discuss in this paper is the tendency of decisions to become worse because of the time pressure of a crisis.

2: Our approach differs from other equilibrium selection theories in that it assigns a probability to each equilibrium. Van Damme (1983) surveys the equilibrium selection problem, and all the theories he cites are deterministic in that they try to choose a single equilibrium or set of equilibria. As the game parameters change, equilibria jump into and out of the set discontinuously. Theories like this are too strong for our purposes, since we recognize that some non-game-theoretical factors will influence the player's judgements about which equilibrium to choose, and abruptly changing an equilibrium from an impossibility to a certainty based only on the utilities leaves no room for these factors. Using such a theory would imply that up to a limit adding warheads would create no danger at all, but beyond that war would be certain.

Our theory can be regarded as a probabilistic generalization of Harsanyi and Selten's equilibrium selection method for 2x2 games with two strong equilibrium (1980, Ch.2), which essentially uses Axioms 2, 3 and 5.

3: One difference between Ellsberg's and our own approaches is that he did not include a fourth outcome representing decisions by both to attack. In our model two sides can see themselves heading to war and decide to attack, but his conception was that a mutual decision to attack was negligibly likely -- an attack occurs by surprise as each side's attitudes and rationality vary over time in ways unknown to the other side. In the analysis here we have stretched his model in our direction by adding a fourth pair of payoffs to his matrix, calculated as the averages of the other two war payoffs as was done for Matrix 2.

Appendix. Proofs of theorems.

The essentials of Lemmas 1 and 2 were first proven by Luce (1959a,b). The first lemma states that the ratio of the probabilities of choosing two outcomes is constant, independent of other outcomes. The set of outcomes O^+ is defined as those where both players receive a positive payoff.

Lemma 1. For any a,b,c in O^+ and any P satisfying Axiom 1 and Axiom 4

$$\frac{P_{\langle abc\rangle}(a)}{P_{\langle abc\rangle}(b)} = \frac{P_{\langle ab\rangle}(a)}{P_{\langle ab\rangle}(b)} .$$

Proof of Lemma 1. Applying the definition of conditional probability,

$$P_{\langle abc\rangle}(a) = P_{\langle abc\rangle}(a|a \text{ or } b)\, P_{\langle ab\rangle}(a \text{ or } b).$$

Expanding $P_{\langle ab\rangle}(a \text{ or } b)$ and using Axiom 4 yields

$$P_{\langle abc\rangle}(a) = P^{\langle ab\rangle}(a)[P^{\langle abc\rangle}(a) + P_{\langle abc\rangle}(b)],$$

and rearranging,

$$\frac{P_{\langle abc\rangle}(a)}{P^{\langle abc\rangle}(b)} = \frac{P_{\langle ab\rangle}(a)}{1-P^{\langle ab\rangle}(a)} = \frac{P^{\langle ab\rangle}(a)}{P_{\langle ab\rangle}(b)} .$$ Q.E.D.

Lemma 2. If P satisfies Axioms 1-4, there exists a real-valued function w such that for any 2x2 unanimity game $\langle ab\rangle$,

$$P_{\langle ab\rangle}(a) = \frac{w(a)}{w(a) + w(b)} ,$$

and for any 3x3 unanimity game $\langle abc\rangle$,

$$P_{\langle abc\rangle}(a) = \frac{w(a)}{w(a) + w(b) + w(c)} .$$

Also, the function w is a ratio scale, i.e., another function w' will generate the correct probabilities if and only if it is proportional to w.

The proof will also show all values of w must have the same sign. We can take the sign to be positive and interpret w(a) as a weight that measures the "attractiveness" of outcome a.

The order of quantifiers in Lemma 2 is important. Given a probability function and any game, one can easily find a function w that works, but Lemma 2 says more: for a given probability function there is a weighting function on the utilities that reproduces the probabilities for all game simultaneously.

Proof of Lemma 2.

By Lemma 1, $P_{\langle abc\rangle}(b) = \frac{P_{\langle ab\rangle}(b)}{P_{\langle ab\rangle}(a)} P_{\langle abc\rangle}(a)$.

Since

$$\frac{P_{\langle abc\rangle}(c)}{P_{\langle abc\rangle}(b)} = \frac{P_{\langle bc\rangle}(c)}{P_{\langle bc\rangle}(b)}, \text{ then } P_{\langle abc\rangle}(c) = \frac{P_{\langle bc\rangle}(c)}{P_{\langle bc\rangle}(b)} \frac{P_{\langle ab\rangle}(b)}{P_{\langle ab\rangle}(a)} P_{\langle abc\rangle}(a).$$

Substituting the two expressions in $P_{\langle abc\rangle}(a) + P_{\langle abc\rangle}(b) + P_{\langle abc\rangle}(c) = 1$ and rearranging terms, gives

$$P_{\langle abc\rangle}(a) = \frac{\frac{P_{\langle bc\rangle}(b)}{P_{\langle bc\rangle}(c)} \frac{P_{\langle ab\rangle}(a)}{P_{\langle ab\rangle}(b)}}{1 + \frac{P_{\langle bc\rangle}(b)}{P_{\langle bc\rangle}(c)} + \frac{P_{\langle bc\rangle}(b)}{P_{\langle bc\rangle}(c)} \frac{P_{\langle ab\rangle}(a)}{P_{\langle ab\rangle}(b)}} . \qquad (1)$$

Choose some e in O^+ and define $w(e)=1$ and $w(a) = \frac{P_{\langle ae\rangle}(a)}{P_{\langle ae\rangle}(e)}$ for a in O^+.

The function w will be well-defined and positive since both numerator and denominator are positive.

By Lemma 1 $\frac{P_{\langle ab\rangle}(a)}{P_{\langle ab\rangle}(b)} = \frac{P_{\langle abe\rangle}(a)}{P_{\langle abe\rangle}(b)}$, but

$$P_{\langle abe\rangle}(a) = P_{\langle abe\rangle}(e) \frac{P_{\langle ae\rangle}(a)}{P_{\langle ae\rangle}(e)} \text{ and } P_{\langle abe\rangle}(b) = P_{\langle abe\rangle}(e) \frac{P_{\langle ae\rangle}(b)}{P_{\langle ae\rangle}(e)},$$

so $\frac{P_{\langle ab\rangle}(a)}{P_{\langle ab\rangle}(b)} = \frac{w(a)}{w(b)}$ holds along with an analogous expression for <bc>.

Substituting the two in (1) gives the formula for the 3x3 case in the lemma. To derive the 2x2 case we use the 3x3 expression plus the following formula derived from writing $P_{\langle abc\rangle}(a)$ as $P_{\langle abc\rangle}(a \text{ or } b)\, P_{\langle ab\rangle}(a)$.

$$P_{\langle ab\rangle}(a) = \frac{P_{\langle abc\rangle}(a)}{P_{\langle abc\rangle}(a) + P_{\langle abc\rangle}(b)} .$$

To show that w is a ratio scale we first show that any change of unit of w will generate the correct probabilities. If $w' = kw$ for some non-zero constant k, clearly w' will yield the two formulae of the theorem. Next we show that all other functions w' satisfying the axioms must be of the form $w' = kw$. Assume there exists a w' that is not, so that $w(a)/w(a) \neq w'(b)/w'(b)$.

Proof of Lemma 2.

By Lemma 1, $P_{\langle abc\rangle}(b) = \dfrac{P_{\langle ab\rangle}(b)}{P_{\langle ab\rangle}(a)} P_{\langle abc\rangle}(a)$.

Since

$$\frac{P_{\langle abc\rangle}(c)}{P_{\langle abc\rangle}(b)} = \frac{P_{\langle bc\rangle}(c)}{P_{\langle bc\rangle}(b)}, \text{ then } P_{\langle abc\rangle}(c) = \frac{P_{\langle bc\rangle}(c)}{P_{\langle bc\rangle}(b)} \frac{P_{\langle ab\rangle}(b)}{P_{\langle ab\rangle}(a)} P_{\langle abc\rangle}(a).$$

Substituting the two expressions in $P_{\langle abc\rangle}(a) + P_{\langle abc\rangle}(b) + P_{\langle abc\rangle}(c) = 1$ and rearranging terms, gives

$$P_{\langle abc\rangle}(a) = \frac{\dfrac{P_{\langle bc\rangle}(b)}{P_{\langle bc\rangle}(c)} \dfrac{P_{\langle ab\rangle}(a)}{P_{\langle ab\rangle}(b)}}{1 + \dfrac{P_{\langle bc\rangle}(b)}{P_{\langle bc\rangle}(c)} + \dfrac{P_{\langle bc\rangle}(b)}{P_{\langle bc\rangle}(c)} \dfrac{P_{\langle ab\rangle}(a)}{P_{\langle ab\rangle}(b)}} . \qquad (1)$$

Choose some e in O^+ and define $w(e)=1$ and $w(a) = \dfrac{P_{\langle ae\rangle}(a)}{P_{\langle ae\rangle}(e)}$ for a in O^+.

The function w will be well-defined and positive since both numerator and denominator are positive.

By Lemma 1 $\dfrac{P_{\langle ab\rangle}(a)}{P_{\langle ab\rangle}(b)} = \dfrac{P_{\langle abe\rangle}(a)}{P_{\langle abe\rangle}(b)}$, but

$$P_{\langle abe\rangle}(a) = P_{\langle abe\rangle}(e) \frac{P_{\langle ae\rangle}(a)}{P_{\langle ae\rangle}(e)} \text{ and } P_{\langle abe\rangle}(b) = P_{\langle abe\rangle}(e) \frac{P_{\langle ae\rangle}(b)}{P_{\langle ae\rangle}(e)},$$

so $\dfrac{P_{\langle ab\rangle}(a)}{P_{\langle ab\rangle}(b)} = \dfrac{w(a)}{w(b)}$ holds along with an analogous expression for <bc>.

Substituting the two in (1) gives the formula for the 3x3 case in the lemma. To derive the 2x2 case we use the 3x3 expression plus the following formula derived from writing $P_{\langle abc\rangle}(a)$ as $P_{\langle abc\rangle}(a \text{ or } b)\, P_{\langle ab\rangle}(a)$.

$$P_{\langle ab\rangle}(a) = \frac{P_{\langle abc\rangle}(a)}{P_{\langle abc\rangle}(a) + P_{\langle abc\rangle}(b)} .$$

To show that w is a ratio scale we first show that any change of unit of w will generate the correct probabilities. If $w' = kw$ for some non-zero constant k, clearly w' will yield the two formulae of the theorem. Next we show that all other functions w' satisfying the axioms must be of the form $w' = kw$. Assume there exists a w' that is not, so that $w(a)/w(a) \neq w'(b)/w'(b)$.

Then $$\frac{w(a)}{w(a) + w(b)} \neq \frac{w'(a)}{w'(a) + w'(b)} .$$

However this is impossible since the two sides of the equation are alternative formulae for $P_{\langle ab \rangle}(a)$. Q.E.D.

Lemma 3. If P satisfies Axioms 1-5, there exists $g > 0$ such that for any 2x2 unanimity game $\langle ab \rangle$

$$P_{\langle ab \rangle}(a) = \frac{(u_a v_a)^g}{(u_a v_a)^g + (u_b v_b)^g} ,$$

and for any 3x3 unanimity game $\langle abc \rangle$

$$P_{\langle abc \rangle}(a) = \frac{(u_a v_a)^g}{(u_a v_a)^g + (u_b v_b)^g + (u_c v_c)^g} .$$

Lemma 3 extends Lemma 2 by giving the form of the weighting function as $w(u,v) = k(uv)^g$, when the probability function satisfies Axiom 5.

The reason why w must be a power function is instructive. (Our discussion here owes much to the exchange between Luce (1959b, 1962) and Rozeboom (1962).) A scientific law typically gives a relation among several physical qualities by stating a functional rule relating their numerical measures. The numerical measures of the qualities are usually not unique, for example, they may be ratio scales so that any other set of numbers proportional to the first set would do as well. (Our theory in particular involves ratio scales. The weighting function w is a ratio scale by Theorem 1, and u and v are ratio scales since they represent the utility increment over the zero-point.) If the unit of measurement of one of the ratio scales changes, some other variable in the law must also change if the numerical functional relationship is to remain true. Since the measures are ratio scales they can change in only one way: multiplication by a positive constant.

As an example of an impossible law suppose two ratio scales x and y are related by $y = e^{-x}$. Changing the unit of the x-scale by setting $x' = 2x$ results in values $y' = e^{-x'} = e^{-2x}$, which are not proportional to the corresponding values of y. In other words, the two laws, $y = e^{-x}$, and $y' = e^{-2x}$ disagree non-trivially about the shape of y's decline with x. The essence of the proof of Theorem 2 is that only for functional laws of the form $y = k\,x^c$ does a proportional change in one variable yield a proportional change in the other.

At first glance many valid laws seem to violate this requirement. For example the law of radioactive decay states $A = A_o\, e^{-d/L}$, with ratio scale variables d for the duration of time the isotope has been decaying, A_o and A for the initial and current amount of the isotope, and L a parameter proportional to the half-life of the isotope. The resolution of the puzzle is simple: a rescaling of d requires a rescaling of L, which is in units of

time. The variable in the exponent has in fact been rendered dimensionless by division by L, and so there is no need to rescale A to maintain the truth of the numerical relationship.

Thus the restriction on the possible functions relating two ratio scales may not apply if the functions contain dimensional parameters. We can be sure that our functions P or w will contain no such parameters since w can depend only on the abstract game, by Axiom 2. Dimensional parameters say something about the real world and must be found by empirical investigation, e.g., the fact that a material has a certain half-life requires observation, but Axiom 2 makes the strong assumption that the probabilities are invariant when the utilities in the game are changed in certain ways, irrespective of any changes in the utilities for outcomes outside the game. In the domain of two-person conflict, the observer might conceivably use an external utility standard such as death or enjoyment of the world's wealth, but Axiom 2 disallows these considerations.

Proof of Lemma 3. Since by Axiom 2 P_G depends only on the utilities of the outcomes, and since by Lemma 2 w(a) depends only on the outcome a, we can express w(a) as $w(u_a,v_a)$. Consider two 2x2 unanimity games <ab> and <a'b'> where the latter is generated from the former by multiplying the first player's utilities by a positive constant k. Using the formula of Lemma 2 for $P_{<ab>}(a)$ and $P_{<a'b'>}(a')$ and rearranging gives

$$w(ku_a,v_a) = [w(ku_b,v_b)/w(u_b,v_b)]\ w(u_a,v_a). \qquad (1)$$

Since the above can be derived for any outcome b, the factor involving b must be a function of k alone, so we may write it $K(k,v_a)$, or $K_v(k)$, where v_a has been written simply as v. Thus from the above formula we have

$$w(ku_a,v) = K_v(k)\ w(u_a,v)$$

Choosing $u_a = 1$ gives $K_v(k) = w(k,v)/w(1,v)$ and the above equation becomes

$$w(ku_a,v) = w(k,v)w(u_a,v)/w(1,v).$$

Let $t_v(.)$ be the function $\log[w(.,v)/w(1,v)]$.

Then $t_v(ku_a) = \log[w(ku_a,v)/w(1,v)]$

$$= \log \frac{w(k,v)\ w(u_a,v)}{w(1,v)\ w(1,v)}$$

$$= t_v(k) + t_v(u_a).$$

The function P is continuous in ku_a, so it can be shown that w and thus t_v are continuous. The above equation for t is an instance of the functional equation $f(xy) = f(x) + f(y)$ and therefore under these continuity conditions it has the unique solution

$$t_v(u) = b(v) \log u \text{ for some b independent of u, so that}$$

$$w(u,v) = c(v)\ u^{b(v)}. \qquad (2)$$

time. The variable in the exponent has in fact been rendered dimensionless by division by L, and so there is no need to rescale A to maintain the truth of the numerical relationship.

Thus the restriction on the possible functions relating two ratio scales may not apply if the functions contain dimensional parameters. We can be sure that our functions P or w will contain no such parameters since w can depend only on the abstract game, by Axiom 2. Dimensional parameters say something about the real world and must be found by empirical investigation, e.g., the fact that a material has a certain half-life requires observation, but Axiom 2 makes the strong assumption that the probabilities are invariant when the utilities in the game are changed in certain ways, irrespective of any changes in the utilities for outcomes outside the game. In the domain of two-person conflict, the observer might conceivably use an external utility standard such as death or enjoyment of the world's wealth, but Axiom 2 disallows these considerations.

Proof of Lemma 3. Since by Axiom 2 P_G depends only on the utilities of the outcomes, and since by Lemma 2 w(a) depends only on the outcome a, we can express w(a) as $w(u_a,v_a)$. Consider two 2x2 unanimity games <ab> and <a'b'> where the latter is generated from the former by multiplying the first player's utilities by a positive constant k. Using the formula of Lemma 2 for $P_{<ab>}(a)$ and $P_{<a'b'>}(a')$ and rearranging gives

$$w(ku_a,v_a) = [w(ku_b,v_b)/w(u_b,v_b)]\ w(u_a,v_a). \qquad (1)$$

Since the above can be derived for any outcome b, the factor involving b must be a function of k alone, so we may write it $K(k,v_a)$, or $K_v(k)$, where v_a has been written simply as v. Thus from the above formula we have

$$w(ku_a,v) = K_v(k)\ w(u_a,v)$$

Choosing $u_a = 1$ gives $K_v(k) = w(k,v)/w(1,v)$ and the above equation becomes

$$w(ku_a,v) = w(k,v)w(u_a,v)/w(1,v).$$

Let $t_v(.)$ be the function $\log[w(.,v)/w(1,v)]$.

Then $t_v(ku_a) = \log[w(ku_a,v)/w(1,v)]$

$$= \log \frac{w(k,v)\ w(u_a,v)}{w(1,v)\ w(1,v)}$$

$$= t_v(k) + t_v(u_a).$$

The function P is continuous in ku_a, so it can be shown that w and thus t_v are continuous. The above equation for t is an instance of the functional equation f(xy) = f(x) + f(y) and therefore under these continuity conditions it has the unique solution

$$t_v(u) = b(v) \log u \text{ for some b independent of u, so that}$$

$$w(u,v) = c(v)\ u^{b(v)}. \qquad (2)$$

By symmetry of the players implied by Axiom 2, $w(u,v) = w(v,u)$, so it is also true that

$$w(u,v) = c(u)\, v^{b(u)}. \qquad (3)$$

Setting $v = 1$ in the two formulae and comparing gives

$$c(u) = c(1)\, v^{b(1)}.$$

Substituting this value of $c(v)$ in the first formula for $w(u,v)$ yields

$$w(u,v) = c(1)\, v^{b(1)}\, u^{b(v)}, \qquad (4)$$

and we can similarly derive

$$w(u,v) = c(1)\, u^{b(1)}\, v^{b(u)}. \qquad (5)$$

Equating (4) and (5), letting $g = b(1)$ and taking logarithms,

$$[b(v) - g]/\log v = [b(u) - g]/\log u.$$

Since u and v can be varied independently both sides must equal a constant d independent of u or v. Therefore $b(u) = d \log u + g$. Substituting this formula in (5) gives

$$w(u,v) = c(1)\, uvg\, v^{d \log u}. \qquad (6)$$

By inserting (6) in (1) we can derive that $d = 0$ and thus that w is a power of the product uv. Using this expression in the formulae in Lemma 2 gives the results. Q.E.D.

Proof of Theorem 1. The critical probability in Matrix 1 for player i is defined as the threshold probability such that at values higher than this player 1 should choose to refrain, and at values lower should choose to attack. Any operation on the payoff matrix that does not change the critical probability will not change player 1's best reply function, and the probability is not changed by adding a constant to the two payoffs of player 1 in a single column. Likewise player 2's critical probability is not altered by adding a constant to 2's payoffs that appear in a single row. Thus by Axiom 5 we can transform one matrix into another by these operations without altering the best reply function, and if it is a unanimity game, use Lemma 3 to calculate the probability function.

By adding a_1 to 1's payoffs in column 1, r_1 to 1's payoffs in column 2, a_2 to 2's payoffs in row 1 and r_2 to 2's payoffs in row 2, Matrix 1 becomes Matrix 4.

a_1, a_2	0, 0
0, 0	$\frac{r_1-a_1}{2}, \frac{r_2-a_2}{2}$

Matrix 4.

Applying Lemma 3 to Matrix 4 gives $P_G(\text{War})$ as in the theorem. Q.E.D.

Proof of Theorem 2. It is easy to verify that the probability of war as given in Theorem 1 can be expressed as $1/[1+(4/CII)^g]$. It follows that all increase with CII and thus all are ordinally equivalent to each other. Q.E.D.

Proof of Theorem 3. The critical risk of player i can calculated to be $2a_i/(a_i+r_i)$ and the critical odds to be $2a_i/(r_i-a_i)$. The expression for CII is seen to be $1/4o_1o_2$ and is thus monotonic with the inverse of the product of the critical odds. Q.E.D.

Applying Lemma 3 to Matrix 4 gives $P_G(\text{War})$ as in the theorem. Q.E.D.

Proof of Theorem 2. It is easy to verify that the probability of war as given in Theorem 1 can be expressed as $1/[1+(4/CII)^g]$. It follows that all increase with CII and thus all are ordinally equivalent to each other. Q.E.D.

Proof of Theorem 3. The critical risk of player i can calculated to be $2a_i/(a_i+r_i)$ and the critical odds to be $2a_i/(r_i-a_i)$. The expression for CII is seen to be $1/4o_1o_2$ and is thus monotonic with the inverse of the product of the critical odds. Q.E.D.